# Vital Signs
# Long-Term Aquatic Monitoring Projects:

# Part B

# Planning Process Steps:
# Issues to Consider And
# Then to Document In

# A Detailed Study Plan That Includes
# A Quality Assurance Project Plan (QAPP)
# And Monitoring "Protocols"
# (Including Standard Operating Procedures)

**January 2, 2004**
**Draft Update of**

**http://science.nature.nps.gov/im/monitor/protocols/wqPartB.doc**

**Roy J. Irwin, WRD, NPS**

**Send Peer Review Comments to:**
Roy_Irwin@NPS.gov

**Citation:**
**Irwin, R.J. (2004). Draft Part B of Aquatic Habitat Park Service guidance for Park Service Vital Signs Monitoring. Planning Process Steps: Issues To Consider And Then to Document In A Detailed Study Plan That Includes A Quality Assurance Project Plan (QAPP) And Monitoring "Protocols" (Standard Operating Procedures), Available on Internet**
**http://science.nature.nps.gov/im/monitor/protocols/wqPartB.doc**

# TABLE OF CONTENTS

**When viewing the following table of contents in Word, Click (IN 2002 or later VERSIONS OF WORD, PRESS THE CONTROL KEY BEFORE CLICKING) on Section Titles and Word Will Take You to the Selected Section:**

## INTRODUCTION:

**"Always do the right thing. This will gratify some people and astonish the rest" Mark Twain**

**This document summarizes a group-consensus thought-process for developing Quality Assurance/Quality Control (QA/QC) and other detailed study planning basics that need to be summarized in a detailed study plan for long-term aquatic monitoring of water quality, sediment quality, and/or aquatic biological (status and trend) "Vital Signs." The same basic steps are recommended for short term or modestly funded projects, but limited budgets typically do not allow one to take each step as far.**

**The standard VS guidance for monitoring plans and for protocols, including proposed numbering systems, is fully consistent with the recommendations herein. It is made clear that quality control and calibration and other method details need to be place in quality control and method standard operating procedures provided for each protocol.**

**The goals of this process recommended herein are to:**

1) **Provide effective pre-project peer-review.**
2) **Select and document study designs that are optimal and "make sense."**
3) **Ensure that the data collected is adequate in terms of quality, quantity, and relevance.**
4) **Ensure that all key parties agree about what is to be done and the reasons for doing it.**
5) **Ensure that the data collected has a good chance of making the transition from raw data to helpful "information" relevant to specific questions and/or important decisions.**
6) **Help the proposed project avoid some of the more common "junk science" pitfalls.**

**It is impossible to suggest QC performance standards, data quality objectives, or detailed standard operating procedure (SOP) protocols that would be optimal for every situation nationwide. One size cannot fit all. Therefore, many of the recommended data quality objectives and QC performance standards listed below are framed as "typical" NPS objectives. This emphasizes the fact that this guidance is more of a suggested thought-process and suggested documentation process rather than a series of absolute requirements.**

**Most other state and federal agencies doing water quality or contaminants monitoring long ago discovered that a QA/QC planning process was necessary to make sure that monitoring information was credible, useful for intended purposes, and defendable. The NPS planning process recommended herein synthesizes key**

**concepts from numerous groups within EPA, the (inter-agency) National Water Quality Monitoring Council, and USGS.**

**The planning steps suggested herein are not only generally consistent with QA/QC guidance from these groups, but also with many other groups (see Appendix A for details).**

**The guidance herein is also designed to be consistent with Department of Interior guidelines to comply with Section 515 of the Treasury and General Government Appropriations Act for FY2001 (Public Law 106-554, http://www.doi.gov/ocio/guidelines/515Guides.pdf and with proposed DOI codes of ethical conduct. The information quality guidelines emphasize Quality Assurance/Quality Control (QA/QC) as well as QA/QC basics such as**

1. **Objective peer review**
2. **Ensuring and maximizing factors such as the "quality, utility, objectivity, and integrity of the information"**
3. **A high degree of transparency about data and methods used to generate data disseminated to the public.**

**Data collected for other agencies must meet data quality guidelines of other agencies. Most of these focus on QA/QC basics such as precision, lack of bias, and transparency. For example, "EPA recognizes that influential scientific, financial, or statistical information should be subject to a higher degree of quality (for example, transparency about data and methods) than information that may not have a clear and substantial impact on important public policies or private sector decisions. A higher degree of transparency about data and methods will facilitate the reproducibility of such information by qualified third parties, to an acceptable degree of imprecision" (http://www.epa.gov/oei/qualityguidelines/EPA_OEI_IQG_FINAL_10-2002.pdf).**

**Guidance herein is consistent with many of the same good science, ethical behavior, and quality assurance concepts utilized by other federal agencies and other monitoring and research organizations (and being considered for inclusion in the DOI Code of Ethical Conduct currently being developed), including:**

1. **Disclosure of methods in enough detail to allow others to attempt to reproduce the information.**
2. **Peer Review and openness to constructive criticism**
3. **Quality and objectivity of information produced**
4. **Adherence to established quality assurance and quality control (QA/QC) programs**

**The relatively non-controversial and easy early fixes of federal and State environmental laws are now mostly behind us. Further improvements tend to be more expensive and to impact potentially greater numbers of responsible parties. In this era, those potentially responsible for environmental impacts have shown a tendency to be quick to bring lawsuits and to otherwise generally challenge the**

**credibility of data that might incriminate them. Thus, reporting credible data has become even more critical than before.**

**To help ensure the success of expensive long-term monitoring, going through a thorough QA/QC and "good science" project planning thought process, as is recommended herein, is not too much to ask of project planners. Even many volunteer monitoring groups are now using planning processes almost this lengthy and rigorous (http://www.epa.gov/volunteer/qappcovr.htm or http://www.epa.gov/volunteer/qapp/vol_qapp.pdf). In 2003 and beyond, professionals in the Park Service should do no less.**

## PHASE 1 FOR THE DEVELOPMENT OF VITAL SIGNS NEWORK MONITORING PLANS

**During Phase 1 of the VS monitoring program, networks should (http://science.nature.nps.gov/im/monitor/approach.htm):**

1. **Form a network Board of Directors and a Science Advisory Committee.**
2. **Summarize existing data and understanding.**
3. **Prepare for and hold a scoping workshop (at least part of this task shall be completed under phase 1** (http://www1.nrintra.nps.gov/im/monitor/VSM_3Phase.doc).
4. **Develop Problem Statements and Identify Values to Be Protected (Step II-A).**
5. **Write a summary of past information and background information (Step II-B).**
6. **Develop Objectives and more detailed questions to be answered (Step II-C).**
7. **Develop Conceptual Models (Step III-A).**

**After the scoping workshop is accomplished, a smaller group (the small-group project planning team of neutral technical experts described below) should gather to discuss and document decisions related to steps in 4-7 above. The small-group project planning team ("the team" described in more detail in section 1-C, below) will ordinarily consist of the Science Advisory Committee plus technical experts (those with more detailed and specialized expertise on practical statistics, study designs, and QA/QC.**

# PLANNING PROCESS OUTLINE:

## Phase 1:

## PRELIMINARY STEPS;

**First networks gather and summarize information:**

**Even in short term or very modestly funded projects, a typical first step is summarizing information content from past data. For long-term monitoring it is even more important to summarize past data to highlight potential issues and trends.**

**Long-term monitoring networks should spend the first year or so determining Park priorities for resources to protect and analyzing past information for quality, usefulness, general information content, potential hints of trends, and patterns of temporal and spatial variability.**

**Information that needs to be gathered in preparation for plan documentation includes, among other things (http://science.nature.nps.gov/im/monitor/monplan.doc):**

1) **The most important natural resources in the network and parks.**
2) **The importance of these resources in a regional or national context**
3) **The most important management issues and scientific issues for each park.**
4) **The most important agents of change and stressors that may cause changes in park resources?**
5) **An overview of monitoring that has occurred in each park in the past.**
6) **Any widely-accepted monitoring efforts used on adjacent lands by other agencies.**

**When looking at past data, keep in mind that much older data is deficient in QA/QC and metadata documentation and is therefore not optimally credible. Metadata is data about data, accompanying information needed to interpret the meaning and quality of the data. Older data sets often lack sufficient metadata and are often not comparable to newer information because of method changes. Older data generated by one agency or group are often not comparable to data sets generated by others because of method differences or QC performance differences.**

**However, the issue is typically not whether or not old data is perfect or optimal, but whether or not it is useful (See definitions of useful quantitative data and useful qualitative data in the definitions section at the end of the appendices). When assessing the usefulness of past data, data users must make an assessment about whether or not the older data are accurate or comparable ENOUGH to use in comparisons, hints about issues or trends, or even estimates of basic variability.**

**When looking at older data, typically one also attempts to assess whether or not there is valuable information content that hints about resource condition versus various potential stressors at sites of concern compared to relatively pristine "control" sites (in the Park or somewhere else in the same ecoregion, watershed, or other relevant ecological unit)? What does past data hint about the ecological relevance of potential variables that might be measured and the degree to which the variables can be used to help discriminate between control sites vs. sites impacted by various specific stressors?**

**The popular generalization that "all old data are bad data" rings true in many cases but also is becoming less true as time goes along and more investigators are beefing up performance-based QA/QC and metadata. See definitions of useful quantitative data in the definitions section at the end of the appendices for additional guidance on assessing the quality and usability of past data.**

**When summarizing potential issues relevant to monitoring, planners should consider sources of pollution or other suspect stressors that are:**

1) **Nearby,**
2) **Upstream, or**
3) **Upwind.**

**Superfund sites, large industrial areas, urban areas, large point sources discharges into water, regional non-point sources considered important, and other suspect regional or local sources should be summarized. EPA databases can be helpful in this task. For example, Envirofacts at (http://www.epa.gov/enviro) provides data which the user can then map, if desired, using Enviromapper. The EnviroMapper StoreFront (http://www.epa.gov/enviro/html/em/index.html) is also a useful site for water, watershed, and other info (including Window to My Environment).**

**Those planning long-term monitoring should consider regional and even national trends. For example, excess nutrients and problems with mercury are major issues worldwide and throughout large parts of the U.S. Most Parks are impacted by nearby habitat changes. PAHs and endocrine disrupters appear to be of increasing concern, while lead, PCBs, DDE/DDT and PCB concerns are often at least slowly decreasing in many parts of the U.S. (See Appendix A for additional details).**

**Identify Context:**

**Once planners have summarized past information and park resource protection priorities, they should identify the proportions of the proposed long-term monitoring that are related to regulatory issues vs. the proportions related to more generalized status and trend monitoring.**

**The monitoring is considered regulatory context if the issue relates to:**

1. **Waters thought to be degraded or impaired, but which are not listed by states (but which possibly could qualify for listing if the States is provided sufficient credible information).**
2. **Waters already officially listed as impaired by States.**
3. **Waters listed as Outstanding Natural Resource Waters (ONRW) by States, or,**
4. **Pristine waters that might be listed as ONRWs if sufficient credible information is presented to the state, or**
5. **Waters of special Anti-Degradation interest, typically not only ONRW areas and pristine waters, but also unimpaired waters in general, or**
6. **Waters utilized by endangered species, or**
7. **Waters impacted by nearby regulated sources such as CERCLA or RCRA sites, or**
8. **Waters of other special interest related to other regulatory or standards compliance issues.**

**The NPS GPRA water goal is goal is that by September 30, 2005, “85% of 265 Park units will have unimpaired water quality.” Since monitoring funding is in short supply in the NPS, hopefully VS and other I&M monitoring will contribute information useful to deciding whether or not GPRA goals are being met. NPS will be moving towards DOI GPRA goals based on the percent of waters meeting EPA approved water quality standards.**

**Therefore, regulatory monitoring related to issues such as those listed above should be the first priority and account for perhaps 60% of monitoring done with WRD-managed (supplemental) vital signs funding.**

**Per the original implementation plan for the water quality portion of Vital Signs monitoring, the other 40% of funding might typically be devoted to general status and trends monitoring of pristine sites, monitoring that might respond tounknown future stressors.**

**For more detail, see separate WRD guidance entitled Identification And Monitoring Of Priority Impaired And Pristine Waters (See Part A at http://science.nature.nps.gov/im/monitor/protocols/wqPartA.doc).**

**Plan As A Team:**

**Since the plan for long term monitoring is important to its success, and since different kinds of expertise are needed to assure optimal pre-project peer review, it is suggested that networks assemble a small-group project planning team (hereafter referred to as "the team").**

**The team should plan project details by moving through the rest of this outline and discussing each item. The goal is to come to consensus agreements on the basics of monitoring design and QA/QC. Meeting notes are documented and used for the basis of the detailed study plan (hereafter referred to as "the plan") that includes a quality assurance project plan (QAPP) and detailed standard operating procedure protocols (SOPs).**

**The team should ordinarily consist of the Science Advisory Committee plus technical experts (those with more detailed and specialized expertise on practical statistics, study designs, and QA/QC. It would typically be helpful to include key members of the NPS monitoring network and other monitoring technical experts, WRD technical advisors, and either regulatory or status and trends monitoring advisors from outside the Park Service. Candidates might typically include USGS monitoring specialists, State monitoring specialists, and other monitoring specialists from agencies like EPA, NOAA, or the Fish and Wildlife Service, depending on who is doing related work in the region. At least one person with considerable applied water quality statistical and study/field survey-design expertise should be included on the team. Other "outside" (independent, other agency) monitoring experts with no conflicts of interest should also be included on the team to the extent practicable.**

**In the case of long-term "vital signs" monitoring, the team should also include key members of the board of directors of the vital signs network, including resource management and/or Park management experts (such as resource knowledgeable Superintendents) that can help specify the resources to be protected and the desired future conditions.**

**Finally, it is suggested the small-group pre-project planning team include other potential users of the data, persons that might provide useful pre-project peer review. For example, if the monitoring is to be general status and trends monitoring using USGS methods, USGS is a potential advisor on the monitoring plan and a potential user of the data, so USGS experts should be involved on the planning team. If the monitoring is to be regulatory context monitoring using State**

**regulatory agency methods, the State is a potential advisor and user of the data, so State experts should be involved on the planning team. If the monitoring is to be in estuarine or marine waters, representative of EPA's marine EMAP program and their State collaborators should be included. NPS funds are limited. Therefore, coordination and collaboration with other groups is often desirable or essential. Accordingly, representatives of other monitoring or regulatory groups that may be candidates for collaboration or coordination should be included in the small group whenever possible. Representatives from academia may be included if appropriate, but keep in mind that many (not all) are not that familiar with monitoring or the kinds of rigorous QA/QC suggested herein, since their main interest is typically research rather than monitoring.**

## Small-Group vs. Post-Project Peer Review

**Does requiring post-project peer review or publishing in peer-reviewed journals ensure project and report quality and thus obviate the need for a Quality Assurance Project Plan (QAPP) or the type of small-group pre-project peer review suggested above? No, most of the time this is a highly incomplete solution. Although a helpful, better than nothing step in many cases, post-study peer review is not a total replacement for a thorough QAPP developed early in the project. Early stage peer review of a proposed study design or proposed STUDY PLAN/QAPP before the project begins (by the small group developing the QAPP at minimum) is often very helpful at preventing problems that cannot be corrected at the end of the project.**

**Several scientific prestigious journals that have studied their peer review process have come to the conclusion that the system is not very successful in preventing bad work from being published. Large percentages of papers containing very large and obvious flaws are recommended for publication. Perhaps even more importantly, post-project peer review is done after the study is already completed, too late to prevent major study design flaws.**

## Outside Peer Review

**However, independent pre-project peer review is typically helpful, so in concert with general I&M Vital Signs Program guidance, in addition to small planning group peer review, the team shall obtain additional pre-project peer review of phases I, II, and III of the proposed plan.**

**All peer reviewers, small group and outside, should comply with generic Whitehouse guidance on peer review, including "(a) peer reviewers be selected primarily on the basis of necessary technical expertise, (b) peer reviewers be expected to disclose to agencies prior technical/policy positions they may have taken on the issues at hand, (c) peer reviewers be expected to disclose to agencies their sources of personal and institutional funding (private or public sector), and (d) peer reviews be conducted in an open and rigorous manner " (http://www.whitehouse.gov/omb/inforeg/oira_review-process.html**

**See appendix A for additional details.**

## Monitoring Plan/Approaches and Methods:

**Before developing a monitoring plan and detailed protocols, it is suggested the team first discuss and then come to consensus decisions on the topics outlined below. The outline has logical splits between overall monitoring plan steps considered quality assurance vs. more detailed topics usually considered quality control and included in protocol standard operating procedures (SOPs):**

### Quality Assurance and General Data Quality Objectives (DQOs, Steps I to IV-C):

**In modern scientific thought, quality assurance is not just a last minute task one does at the end of planning, but includes the entire planning process, including carefully thinking through the questions that need to be answered after summarizing what is already known, making sure the data collected are relevant, representative, comparable, and of adequate quality and quantity, and making sure the study design is defendable and "makes sense."**

# I. INTRODUCTION AND BACKGROUND:

**The following steps are related to developing the introduction and background sections of the detailed study plan. They are discussed in more detail in a stepwise process as follows, to help make sure quality assurance "basics" are adequately covered.**

### I-A. Background:

**Background information needed in the planning process and in the plan should typically include not only summaries of past information and other information such as water quality standards gathered in step one, but also other types of information as listed below.**

**In accordance with general I&M vital signs guidance, the introduction and background section of a detailed study plan should summarize:**

- **The purpose of the monitoring program, including a summary of legislation, NPS policy and guidance, Servicewide and network-specific goals for monitoring, Servicewide and park-specific strategic goals for performance management that are relevant to the monitoring, and any statements from park enabling legislation that establish the need to monitor natural resources. Answer the question, "who is interested in the information provided by monitoring, and why?"**

- **Give an overview of each park and its natural resources. More detailed descriptions of each park and its resources could be included in an appendix. What is the importance of the park's natural resources in a regional or national**

**context? For water quality monitoring, identify parks that have waters where constituents exceed water quality standards and are listed on state Clean Water Act 303d lists or constituents that may be threatened to become degraded. Also identify parks that have waters designated as Outstanding National Resource Waters or other special protective designations in their state water quality standards. Draft guidance for identifying these waters is contained in the Vital Signs Monitoring Desk Reference.**

- **What are the most important management issues and scientific issues for each park? What are the most important agents of change and stressors that may cause changes in park resources?**

- **Give an overview of natural resource monitoring that is currently being done in each park or that occurred previously. Describe any widely accepted monitoring efforts used in the general region by other agencies that provide opportunities for data comparability (putting the network's data in context and assisting in interpretation of data collected in parks).**

- **Introduce the overall process used to determine the goals and specific measurable objectives for the monitoring program, and to select the vital signs for monitoring park resources and providing the information needed to manage the parks. More details should be given in section IV (Vital Signs, below).**

**For details see Appendix I-A.**

**Typical NPS quality assurance objectives:**

**The plan should detail relevant park history, relevant legislation, a discussion of the severity of the resource threat, and the problem or needed action.**

**The plan should provide a discussion of the status of any environmental planning, compliance, or permitting processes or documents that have been completed, including National Environmental Policy Act categorical exclusions, Endangered Species Act Section Seven Consultation, NPS planning documents related to water resources, and appropriate cultural and historical clearances.**

**To optimize study designs, any available (and useful) past information on concentrations, data ranges, spatial variability, temporal variability, chemical-parameter water quality standard exceedances (percentage of exceedances, frequency, and timing), documentations of impairment events related to State water quality standard biocriteria, and important response-variable relationships, and all other important information gathered in step I should be summarized in the background section of the plan.**
**More details on how to accomplish this step may be found in Appendix I-A.**

## I-B. Problem Statement/Value(s) To Be Protected:

**In addition to summarizing the general "purpose" of the monitoring program it is also desirable to specifically identify the values to be protected. The reasons the Park Service desires to monitor aquatic habitats usually have something to do with protecting and managing resources. Defining what those resources are, in the context of desired future conditions, is helpful in focusing the overall design of the monitoring plan.**

**The concept of determining desired future conditions is commonly employed in land management and in studies of ecological trends. The desirability of establishing desired future condition goals has been mentioned in NPS I&M monitoring guidance, and various Parks and networks have recognized the need to do monitoring to decide whether or not trends are going in the right directions and whether or not intervention is needed to achieve management objectives.**

**Whether explicitly stated or not, deciding whether or not trends are going in the right direction implies that one has somehow determined optimal or desired future conditions. As mentioned in I&M monitoring guidance "The NPS recognizes the importance of collecting data in a scientifically credible manner so that they can be used to address current and future management issues" (http://science.nature.nps.gov/im/monitor/Indicators).**

**Sometimes desired future conditions have not been determined and initial monitoring phases need to be planned with the specific goal of determining desired future conditions. Many other organizations specifically recognize the need to use monitoring to determine desired future conditions. For example, the Watershed Management Council states "It is important to collect data from undisturbed sites to determine what background conditions are for the water quality parameters of interest. This information can then be used to determine desired future conditions, and reference variability for specific water quality parameters" (http://watershed.org/news/sum_96/monitor.html).**

**The most efficient way to design monitoring is to state what desired future conditions are at the outset. This helps determine "if-then" decision rules discussed in more section IV-C.**

**The draft (out for public review) Department of Interior strategic plan for the Government Performance and Results Act (GPRA) mentions the need to establish desired future conditions in "management plans." Outcome "measures" suggested as examples include (http://www.doi.gov/gpra/stratplan_2_14_2003.html):**

1) **In marine, coastal, wetland, riparian, and upland environments, the percentage of acres achieving desired conditions as specified in management plans,**
2) **For surface waters, the percentage of surface waters that meet EPA water quality criteria.**
3) **For biological resources, the percentage of species of management concern that are managed to self sustaining levels. Note from Roy Irwin, Sustainability has become such a cornerstone concept that there is now a**

**virtual journal on the subject (Virtual Journal of ENVIRONMENTAL SUSTAINABILITY website at: www.elsevier.com/vj/sustainability.**

4) **For threatened or endangered species, the percentage of species listed a decade or more that are stabilized or improved.**

**Likewise, the Park Service strategic plan (GPRA) goal for water is that by September 30, 2005 "85% of 265 Park units will have unimpaired water quality" (http://www.nps.gov/performance/StrategicPlan01-05.pdf).**

**If Park Service management is to be held to these type of performance goals, then "desired future conditions" probably need to be stated more explicitly in future resource management plans and water resources management plans. Likewise, consideration needs to be given to making one goal of future monitoring programs (including Vital Signs) the goal of helping produce the kind of status and trend information required to determine whether or not the trends are towards or away from optimal or desired future conditions.**

**Typical NPS quality assurance and data quality objectives:**

> **The environmental resources (for example: water quality or populations of aquatic life varying up and down within natural or optimal ranges) or other values to be protected should be explicitly identified in the plan. These are typically resources that are highly valued, whether or not they are already considered susceptible or currently at risk. Desired future outcomes, and perceived or potential "problems," (if any) should also be identified.**
>
> **In Park Service habitats already classified as Outstanding Natural Resource Waters (ONRWs), or already known to be pristine but not yet classified as ONRWs, the value to be protected/desired future condition is typically to preserve pristine (un-impacted) condition status. In this case, the goal is typically to prevent degradation and the desired future conditions, including variability ranges during various regional climatic and flow conditions, should be defined as quantitatively as possible in the plan.**

**For more details, see appendix I-B and more generic NPS I&M Internet references, including: http://science.nature.nps.gov/im/monitor/#Indicators and Dale and Beyer, 2001, Challenges in the development and use of ecological indicators (http://science.nature.nps.gov/im/monitor/EcolIndDev.pdf).**

## I-C. Questions to Be Answered/Objectives:

**One of the most important things that study planners can and should do to help optimize study designs is to state clear and very specific questions to be answered, in addition to stating broad objectives. Plain-language (overview)**

**monitoring goals and objectives are often helpful and both network and service-wide (GPRA) goals should be listed, but it is easier to design monitoring to answer specific questions than broad objectives.**

**The specific questions to be answered should have stated boundaries in time and space. Specific questions tend to drive details of study designs and statistics to be used.**

**It is important to avoid so-called type III errors, accurate (low uncertainty) answers to the wrong questions.**

**Will the project answer regulatory questions such as water quality impairment questions (relevant to GPRA goals)? For information on GPRA, see** http://www.nps.gov/performance.

**A study of flawed/failed monitoring projects revealed that many problems could have been avoided if the monitoring had been motivated by a desire to answer a relevant question rather than monitoring for monitoring's sake [L.M. Reid. 2001, The epidemiology of monitoring. Jour. Amer. Water Resources Assn. 37(4): 815-819].**

**A basic generic question, one that can often serve as a starting point for developing more detailed questions is the following: "Is parameter or metric X varying within pristine or un-impacted (desired condition) ranges during various natural conditions and in specific areas of time and space?"**

**After basic questions are formulated, they should be made as specific as possible. For example, at first attempt, one might specify the general objective of the project to be "to determine mercury concentrations in water." A second attempt one might produce the following (still too general) question: "Do the concentrations of mercury in water exceed the state water quality standard?" A final attempt to provide more specificity to the question in time, space, and needed statistics might produce the following much more specific questions:**

> **"If mercury concentrations exceed state water quality standards by x, what is the probability that the sampling will detect the problem?**
>
> **Does the measurement bias-adjusted upper 95% confidence limit based on a t distribution and a minimum sample size of 30, for samples only collected during morning hours, exceed the standard?"**
>
> **"Does the one-hour average concentration (based on a minimum of 5 samples per hour collected at least three times a month for one year, or other minimum sampling specified by the State) of mercury from depth-composited water column samples at randomly chosen sites in a specified reach of river ever exceed the State water quality standard Criteria Maximum Concentration (CMC)?"**
>
> **"What are the percentages of water column mercury concentrations (based on depth-composited water column samples at randomly chosen sites in a specified reach) that exceed state water quality standards, based on the minimum number of samples specified by the State [to conform to the**

> **decision criteria for listing on a 303(d)---impaired waters]? What is the 95% confidence interval on the proportion? Does the percentage or the upper 95% confidence level exceed the minimum percentage required for definition as impaired or partially impaired? Are other listing criteria (fishing closures due to excess mercury in fish, etc.) being exceeded, based on State and/or EPA guidance?"**

**The following is an example of a very general status and trends monitoring question:**

> **"Are the concentrations of mercury in water increasing?"**

**A much more specific question, one that would be much more useful in determining the necessary monitoring design and necessary statistics to be used might be the following:**

> **"Do the results of a properly-framed, nonparametric seasonal Kendall (Mann-Kendall) test for trend detection in water column mercury concentrations reveal a statistically significant trend (with 99% statistical power), based on a minimum10 year data set of monitoring at least four times each 60 day period with intervals between the four samples of not less than 30 hours?"**

**Typical NPS quality assurance and data quality objectives:**

> **Detailed questions to be answered by monitoring should be explicitly identified in the plan. The questions should be as specific as possible in terms of spatial and temporal boundaries, as well as statistics.**

**More detail on how to accomplish the important task of framing specific study questions may be found in Appendix I-C.**

## II. CONCEPTUAL ECOLOGICAL MODELS

### II-A. General Conceptual Ecological Models

**In accordance with overall I&M guidance, the plan should provide a summary of the understanding of the park ecosystem, including conceptual models of relevant ecosystem components developed during the scoping and review process. This summary should focus on aspects of the ecosystem that are relevant to the monitoring program. Guidance and examples of conceptual models can be found at http://science.nature.nps.gov/im/monitor.**

**A key but to often not well documented issue is summarizing what is known about the relationship between the objects to be measured and the value(s) to be protected (or the desired future outcomes).**

**Model understandings typically drive questions and therefore it is often clear that there is a relationship between "the questions to be answered" (the topic covered in the section above) and current conceptual model understandings. As our model understandings change, so do our questions. For example, in coastal ecosystems, a typical older question to be answered, based on relatively simple limnology models was "How does anthropogenic nutrient enrichment cause change in the structure or function of near-shore coastal ecosystems?" As our model understandings have changed, numerous more detailed questions to be answered have evolved, such as "How does nutrient enrichment interact with other stressors (toxic contaminants, fishing harvest, aquaculture, non-indigenous species, habitat loss, climate change, hydrologic manipulations) to change coastal ecosystems?" (J. E. Cloern, 2001, Our evolving conceptual model of the coastal eutrophication problem. Marine Ecology Progress Series 210: 223-253, http://www.int-res.com/articles/meps/210/m210p223.pdf).**

**A potential problem with stressor-driven conceptual models is that they don't all necessarily define desired future conditions. However, in some cases one could perhaps define desired future condition via indicator attributes, as long as those attributes are things like natural water quality or natural populations of invertebrates, or natural patterns of salinity. This only works well if the attributes are values that managers care about protecting, rather than just groups of measures.**

**Conceptual models such as the model for the Everglades (http://science.nature.nps.gov/im/monitor/CERP2.pdf) summarize the current understanding of linkages between stressors and various parts of the ecosystem. Many model diagrams are like those of the NPS Heartland VS monitoring network, models that the tend to "follow a top-to-bottom hierarchy that identifies natural and anthropogenic drivers at the top, then move down through specific stressors, ecological effects resulting from stressors, recommended ecological attributes/indicators, and (at the bottom) measures for each attribute."**

**Such models seem to start with the notion that stressors are will understood, and they do not tend to include the notion of desired future conditions.**

**However, usually the stressors and responses to stressors are not well understood. EPA has recently published extensive guidance on stressor identification (EPA. 2000. Stressor Identification Guidance, EPA-822-B-00-025, available at http://www.epa.gov/ost/biocriteria/stressors/stressorid.pdf).**

**A simple way to include desired future condition in the conceptual model would be to include them as attributes (optimal populations of various aquatic organisms, for example). For other ways, see appendix II-A).**

**Is ecological health a valid metaphor? Planners should be aware that There are logical and statistical pitfalls in using many common "ecosystem health" or "environmental health" or "ecological integrity" or "biological health" or "biological integrity" or similar metaphors, particularly when they involve somewhat arbitrarily combining various metrics into a single index (see appendix II for more detail).**

## II-B. Aquatic Models for Environmental Decision Making

**Although the stressor-based conceptual ecosystem models discussed above are useful, it is also helpful to model or at least think through environmental decision making models, especially when the monitoring is being done at least partly for regulatory reasons.**

**One simple example of such a model has only four boxes (M.A. Harwell et al. 1990. Characterizing Ecosystem Responses to Stress. National Academy of Sciences Press, http://books.nap.edu/books/0309042933/html/94.html#pagetop):**

**1) A "regulatory endpoints" box (endpoints specified by legislation, regulations, and courts). These regulatory endpoints are "translated into" and connected by arrow to**
**2) The endpoints in the next box, the "ecologically meaningful endpoints" box. Each of the ecologically meaningful endpoints is linked to a companion endpoint in the next box. This ecologically meaningful endpoints box is linked by arrow to the next box, the**
**3) The "Indicators of Ecological Effects" Box. The suite of indicators selected in this box "provides the basis for evaluating ecological responses. An arrow leads from this box to:**
**4) A "Monitoring/Characterizing Ecosystems to Stress Box." Once ecosystem health is characterized in this box, and final arrow leads back the regulatory box (box 1, above) for management response and/or regulatory action.**

**This type of model is helpful in planning monitoring in that it forces one to make sure that the indicators chosen in box three relate to changes in ecosystem health in box four. This model also fits the model of how water quality is actually protected by both EPA and the states under the Clean Water Act and other laws.**

**In the model above and the underlying thought process behind implementation of the Clean Water Act, the basis or starting point is protecting ecological integrity and aquatic health. It is assumed that populations and communities of invertebrates, fish, and plants in the healthy aquatic environment optimally vary within normal ranges. These ranges are typically determined at relatively pristine or un-impacted "control" sites in the same ecoregion. It is further assumed that vital signs like pH are varying up an down in normal ranges (6 to 9.5 for pH, for example) in healthy ecosystems. If a healthy population is not present, one assumes a stressor is present and performs a stressor identification process (EPA. 2000. Stressor Identification Guidance, EPA-822-B-00-025, available at http://www.epa.gov/ost/biocriteria/stressors/stressorid.pdf) to determine the stressor(s) most likely to be responsible.**

**The most simple conceptual models would have desired future conditions at the top, various known and suspected stressors in the middle, then vital signs and other indicators on the next row. The next box might include a stressor identification process.**

**Monitoring of the indicators would follow to see if the desired conditions are being restored. None of this should be construed as negating the need for the more general ecological linkages stress-based model explained in step II-A above. Instead, all of these steps are helpful in planning optimal monitoring.**

**If all one has done is the general ecological conceptual model explained in step II-A, it is still also helpful to model or think through in more detail, the relationship and sensitivity uncertainty for the last two lines of the general conceptual models. In this effort one would look more closely at the relationship of the measures (last line) to the natural resources to be protected or the desired future condition (often the attributes line second from bottom). Usually, natural processes or features defined as attributes (such as natural populations of invertebrates varying within natural ranges or natural patterns of salinity) would be part of resources to be protected or desired future conditions.**

**Like endangered species, rare and especially valuable or unique resources in National Parks are rare and special and perhaps deserve "special" protection. "Special protection" could be protection at the individual level (we don't even want to lose one) rather than the population level, and could involve more stringent than normal criteria or than normal state standards for levels of contaminants in water, prey tissue, sediments, or soils.**

**In recognizing the intuitively "special" status of National Park resources, EPA risk assessment guidance pointed out that "Risk managers are more willing to use a risk assessment for making decisions when it is based on ecological values that people care about. Thus, candidates for assessment endpoints include endangered species and aesthetic values such as CLEAN AIR IN NATIONAL PARKS" (EPA 1998: Guidelines for Ecological Risk Assessment Risk Assessment Forum, U.S. Environmental Protection Agency Washington, DC, EPA/630/R-95/002F, April 1998, Final, www.epa.gov/ncea/ecorsk.htm).**

**For more information, see appendix II-B..**

### II-C. Sensitivity of Values to Be Protected to Changes in Model Components, Including Measures (Potential Vital Signs).

**Perhaps one of the most basic kinds of uncertainty associated with conceptual models is the uncertainty associated with the arrows that connect boxes in the diagrams of the conceptual model. Sometimes the arrows that connect boxes have to go through several other boxes before connecting a proposed vital sign measurement to a desired future condition. So another relevant question is what is the overall uncertainty between the proposed vital sign to be measured and the resource to be protected, or the "desired future condition." Optimal vital signs will have a direct or at least traceable connection between the two, and the uncertainty of the connection will be low. Some sort of uncertainty estimate needs to be made between the vital sign variables proposed for measurement and the resources to be protected, or the desired future condition.**

**Quantifying the uncertainty in models can be complex. At a walk before you run stage, one could attempt qualitative "expert judgments" (Type B uncertainty) rankings of say high, medium and low uncertainty.**

looking at the last two levels of some conceptual models, one can often find populations one is trying to protect on the next to the last (attribute) level, and what one is measuring on the lowest level (measures). To what extent will changes in vital signs be used to predict what the change in the population (to be protected)?

Those with the inclination and statistical savvy could go beyond qualitative estimates but some of the methods for doing so are complex (for more information see appendix II-C.

Sensitivity is the amount an estimated measurement or observation changes in response to a true change (in a magnitude one cares about) in an environmental parameter or value of concern. For more information on sensitivity in general, see section V-B below.

In this document, sensitivity is discussed for three levels of organization or concern:

1. Measurement sensitivity, pertaining even to a single value or data point is discussed in section V-B.2.
2. Statistical/study design sensitivity to detect a change is discussed in the section V-B.7.
3. The sensitivity of a proposed vital sign to some other broader value or concept (for example, values to be protected, desired future condition or aquatic ecological integrity). This is the broad type of relationship or model sensitivity being discussed in this section.

In this last type of sensitivity, typically what one is looking for is the ratio of a change in measurement or observation value to a change in the broader value. Is the ratio close to unity or is there a good correlation?

To help make sure the right things are sampled, it is useful to identify the relationship between any initially proposed measurements to be made (what is measured or recorded) and the questions to be answered, as well as and the value(s) to be protected (or the desired future outcome). Since it is often difficult to estimate sensitivity as a ratio or correlation coefficient, it is typically still helpful to qualitatively classify the uncertainty as low, medium, or high.

Would this measure or metric be an important one to include in a biological integrity index or a multivariate analyses of important measures intended give an estimate of overall condition of the resource to be protected?

Why model or at least think through the relationships between vital signs picked and values to be protected or desired future conditions? A study of flawed/failed monitoring projects revealed that many problems could have been avoided if the projects had not picked measurements (often easy or past measurements) that were irrelevant to the problem the projects were intended to address (L.M. Reid. 2001, The epidemiology of monitoring. Jour. Amer. Water Resources Assn. 37(4): 815-819). For more technical information on how to make these determinations, see NPS guidance at (http://science.nature.nps.gov/im/monitor/epacrit.htm#phase1 and EPA ecological indicator guidance at http://www.epa.gov/emap/html/pubs/docs/resdocs/ecol_ind.pdf).

**EPA suggests that when picking indicators, one should consider, "Discriminatory Ability, The ability of the indicator to discriminate differences among sites along a known condition gradient should be critically examined" (http://science.nature.nps.gov/im/monitor/epacrit.htm#phase3). The condition gradient could be something general or integrative, such as eutrophication or ecosystem well-being.**

**Some values have little correlation sensitivity to other values. Other things being equal, it is good pick indicators with a high signal (value) to estimated variability ratio and did not exhibit large, naturally occurring variability.**

**Again, related but different concepts on sensitivity in general are discussed later in section V-B. As discussed in more detail that section, signal to noise ratios are frequently key concepts in measurement sensitivity and one type of low level sensitivity, detection limits. To avoid confusion, the word "noise" should be used only in the context of measurement imprecision (random up and down fluctuations due to imperfections in the measurement process including imprecision of the measuring system and various interferences/disturbances) to differentiate this concept from a related concept being discussed in this section, our best estimates of the sensitivity of a vital sign to larger values of interest, considering true heterogeneity/variability in both the indicator and the more general value of interest (such as ecological integrity).**

**Some might point out that many meters actually measure one thing and give out a read-out in another. For example, field probes do not measure pH directly but instead measure potential in millivolts (mV). In this case the correlation between mV and pH is good enough that the a reliable algorithm can be built into the meter and the answer on the meter read-out is given as pH units rather than mV. However, the concept of trying to relate vital signs to changes in values in protected is different. The correlations are not so good, and the concept being discussed is not simply using one value as a surrogate to calculate value another with a very high degree of correlation. Instead the concept being discussed in this section is response sensitivity of one parameter to another broader value concept (such as eutrophication or biological integrity).**

**In some ways, the sensitivity of a measured parameter to a larger value is more related to model uncertainty (Uncertainty regarding gaps in scientific theory required to make predictions on the basis of causal inferences, see section IX-F.3 for more detail).**

**One Vital Signs network used a step by step process for addressing this type of uncertainty and related factors:**

1. **Is the variable sensitive to stresses on the resource of interest, while showing limited and documented sensitivity to other factors?**
2. **Are there data showing that the variable responds to stress in a predictable manner?**
3. **Does the variable show low background variability?**
4. **Is it feasible possible to monitor these variables in a scientifically sound way on a regional scale? (p. 67 of report at http://pawcatuck.edc.uri.edu/PhaseII/NCBN_PhaseII_Plus_Appendices.zip).**

**An example of a relationship between a parameter that probably has a poor sensitivity to the higher value to be protected would be the following:**

> **Proposed Vital Sign: Different observer's best estimates of "percent embededness in cobbles in stream bottom sediments" based on qualitative habitat observations.**
>
> **Desired Future Condition: A restored, healthy population of topminnows of the genus Fundulus. The sensitivity of the proposed vital sign to the desired future condition would be high, and one might also say uncertainty in the conceptual model relationship between these two variables is high or at least unknown (see section IV-F.3 for more details).**

## Typical NPS Quality Assurance Objectives:

**The relationship between the measurements or observations to be made and the value(s) to be protected (or the desired future outcome) should be identified in the plan. When practical, quantitative estimates of the strength of the relationship should be made as parametric or nonparametric correlation coefficients, statistical uncertainty, or ratios of the in change in vital sign magnitude divided by the change in the broader value of interest (such as the value(s) to be protected or the desired future outcome).**

**If quantitative estimates cannot be made, the uncertainty in the relationship should be characterized as high, medium or low,**

**Unless otherwise justified, the variables and metrics chosen for monitoring should be clearly relevant to desired future outcomes. In other words, the indicator to be monitored should be conceptually relevant to the assessment question (management concern) and to the ecological resource or function at risk. Typically this means that changes in variables or metrics being measured either directly or indirectly contribute to changes in the status of the resources being protected.**

**See appendix II-C for additional details.**

### II-D. Peer Review of Phase I

**The description of this peer review step is placed here to help networks understand how this VS program-required peer review step fits sequentially into the other general and QA/QC monitoring planning steps.**

**In concert with general (not just aquatic habitat) I&M vital signs guidance, the Phase 1 report should 1) describe the formation of the network Board of Directors and science/technical committees, 2) describe the results of the work involved in summarizing existing data and understanding of the park ecosystems, describe the goals and objectives for the monitoring, 3) describe draft conceptual models, and 4) summarize other background work that should be done before the**

**initial selection of vital signs. Most of the background material and development of conceptual models described in Chapters II and III should be written up in a Phase 1 report. The phase 1 report shall be peer-reviewed and approved by the regional I&M coordinator before the network moves on the phase two and selects and prioritizes its vital signs. The Phase 1 report may not include some material that may not be developed until later in the design process, such as specific, measurable objectives and threshold values or "trigger points".**

**The Phase 1 report may include some early thoughts or initial recommendations but not necessarily the final decisions on issues that may be more fully developed later in the design process, such as specific, measurable objectives and threshold values or "trigger points").**

### PHASE 2 FOR THE DEVELOPMENT OF VITAL SIGNS NEWORK MONITORING PLANS:

**Now that the first three steps of NPS VS general (not just aquatic) guidance (http://science.nature.nps.gov/im/monitor/approach.htm) have been completed, phase 2 includes the following steps:**

### IMPLEMENTATION PLAN/APPROACH AND METHOD STEPS IV TO X1:

## III. VITAL SIGNS

### III-A. Selection of Vital Signs

**Before the sampling plan is finalized, it is critical to identify exactly what "vital signs" and other related parameters (if any) are to be measured.**

**Networks that have taken expert opinion polls at workshops or by using Internet polling techniques will often end up with a list of indicators that is longer than the monitoring budget would allow. The list might also be biased by the specialized interests of the particular group of experts included in workshops or polling.**

**Sometimes the excess cost of monitoring the things suggested will be the indirect result of excess true variability (sample heterogeneity) at pristine sites and/or excess measurement uncertainty. Excess measurement uncertainty often results from lack of reproducibility precision and/or excess or unknown measurement systematic error (bias). Potential vital signs with these characteristics are not advisable because these characteristics often make detecting changes or trends impractical or impossible, or (at minimum) tend to drive up the number of samples required after power analyses are done.**

**Therefore, once all the polling (like the Delphi process explained at http://eies.njit.edu/~turoff/Papers/delphi3.html#Introduction) tasks and workshops are finished, a smaller group of unbiased and study design-savvy and statistics-savvy experts typically needs to put potential vital signs through a series of winnowing down steps using stated, fair, and (whenever possible) neutral quantitative criteria.**

# Typical NPS Quality Assurance and Data Quality Objectives:

**In the plan shall detail what vital signs and other parameters will be measured in each of the following three categories:**

## "Required" Freshwater Parameters

**Four water-column parameters are required to be measured at all aquatic long term monitoring sites.**

**For freshwater, they are specific conductance (differs from conductivity by being temperature corrected), dissolved oxygen, pH, and water temperature. In addition, at least a qualitative assessment of flow or water level is suggested, along with photographic documentation of the collection site (a minimum record of one digital site photo).**

**These parameters are often necessary to interpret or calculate other values. These parameters are required partly so that the NPS has some consistent data nationwide (see Part C of this guidance at http://science.nature.nps.gov/im/monitor/protocols/wqPartC.doc for more detail).**

**These parameters are required not so much because they are always more important than other things (like quantitative flow estimates or like biological populations), but because they are basic, relate to the vital signs concept, are easy and relatively inexpensive to measure while already in the field, are also considered "core parameters" by the National Academy of Science's National Research Council (http://books.nap.edu/books/0309075793/html/26.html#pagetop) and by interagency groups.**

**The qualitative assessment of flow or water level typically involves choosing a category from the following options (% of bank full):**

- **Low - (<25% of bank full or at/near base flow)**
- **Intermediate - (25% ≤ Q ≤ 75% bank full)**
- **High - (> 75% bank full)**
- **Over Bank/Flood Stage - (> 100% bank full)**

**If the state in question has a protocol for qualitatively estimating flow, that may be used instead. For example, the State of Texas method is on a Website (http://www.tnrcc.state.tx.us/admin/topdoc/gi/252/swqmproc.pdf). Such methods may also be used in other areas if considered superior, but data produced by qualitative methods should be clearly designated as "qualitative flow estimates."**

**If the waterbody is dry, other water column parameters like pH and conductivity cannot be taken, but recording the fact that the habitat is dry may be important to tracking changes in frequencies of flow or water level conditions. Changing stream flows was singled out as an especially important indicator in the**

**Heinz report on the state of nation's ecosystems (http://www.heinzctr.org/ecosystems/fr_water/indicators.shtml).**

**Where there is a practical way to get quantitative flow without excess cost, quantitative flow is preferred (but not required), usually from one of the following options:**

- **Gage station (e.g. USGS gage at or near site)**
- **Install staff gage and develop rating curve**
- **Manual flow measurement (various methods)**

**There is a need to monitor a core set of parameters nationwide in a consistent manner to have some nationwide comparable data. The four parameters required in both freshwater and marine sites often relate to water quality standards, can be monitored relatively easily in consistently the same way at all networks, and are often needed as normalizing data or metadata useful in the interpretation of other variables. STORET provides separate metadata fill-in-the-blank boxes for temperature, depth, time of day, and day of year. These are recorded by most monitoring programs each time that any water quality collections are made, but many would consider these parameters as routinely reported metadata rather than "required parameters." So in one sense, there are only 3 required "result parameters" since new STORET seems to consider temperature (like depth) as metadata.**

## "Required" Marine and Estuarine Parameters

**The NPS Marine/Estuarine Required-Parameter Working Group (convened in Narragansett, RI, on April 3 and 4, 2002) agreed with the freshwater group to require four basic parameters in marine or estuarine environments (C. Roman, R. Irwin, R. Curry, M. Kolipinski, J. Portnoy, L. Cameron. 2003. White-Paper Report of the Park Service Vital Signs Workgroup for Monitoring Marine and Estuarine Environments. Workgroup Convened April 3-4, 2002, North Atlantic Coast CESU at the University of Rhode Island, Narragansett, RI. (http://science.nature.nps.gov/im/monitor/COREparamMarine.doc). The following is a brief summary (see full paper at the Internet URL cited above for details):**

**The following core water quality parameters were selected for monitoring within marine/estuarine waters of parks:**

**1) Water Temperature (degrees C°), rounded to nearest degree or to the nearest tenth of a degree if justified (see appendices for additional discussions of rounding rules)**

**2) Dissolved Oxygen (mg/L, ordinarily round to two significant figures, usually one whole number and one decimal place, unless otherwise justified)**

**3) pH (pH units, ordinarily round to nearest 0.1 units)**

4) Ionic strength expressed as conductivity and as salinity. Salinity is a value calculated from conductivity, temperature, and if applicable, other factors such as depth or pressure. Unless otherwise justified, it is recommended that salinity should be calculated from conductivity using the equation from the Practical Salinity Scale of 1978. A calculator recommended for this conversation is the UNESCO/John Hopkins U. calculator (http://ioc.unesco.org/oceanteacher/resourcekit/M3/Converters/SeaWaterEquationOfState/Sea%20Water%20Equation%20of%20State%20Calculator.htm.

Except for the recommendation that salinity should be calculated from conductivity rather than calculating specific conductance from conductivity, the four parameters are the same ones chosen by the freshwater working group (http://science.nature.nps.gov/im/monitor/COREparam.doc).

These four parameters represent our recommended essential minimum suite of parameters that should be included in a water quality monitoring. Many parks will and should augment these basic core parameters with additional ones with the objective of implementing a more comprehensive water quality monitoring program.

In what may be perceived as an irony by some, all 4 of these “required parameters” vary diurnally and often seasonally, particularly in estuarine sites, so trend detection is difficult unless one has high quality continuous monitoring data.

In trend detection, data comparability issues require especially strong scrutiny. For example, use of different meters or standard operating procedures (how long one lets a sample settle in the lab before taking a conductivity reading, for example) can cause changes in readings that might wrongly be attributed to trends.

Although trend detection may be complex, many would nevertheless argue that all four of the required parameters are needed for other reasons, including the fact that several of these are often needed as input data needed to calculate other values. Often the required parameters are also needed for site classification purposes.

Other Associated Required Information:

A) Location standard coordinates [for example, the Universal Transverse Mercator (UTM) grid; on USGS quad];
B) Local time (indicating standard or daylight-saving time);
C) Water depth and sample depth;
D) Tidal stage (e.g. high, low, or mid-tide) and direction (ebb, flood or slack water),
E) Estimated Wave Height.
F) Flushing time

G) Tidal range
H) Habitat description

In association with monitoring of the core water quality parameters, the following information should be collected in conjunction with the water quality collection. These data are important for interpreting any observed trends in water quality.

*Meteorological Data*

- Precipitation - ongoing and recent trends.
- Air temperature
- Wind speed and direction
- Barometric pressure

Suggested for optional consideration are those parameters included those recommended by the National Coastal Assessment (NCA). The NCA includes our four core parameters plus Photosynthetically Active Radiation (PAR), water column dissolved nutrients, chlorophyll *a*, and total suspended solids, sediment contaminants, sediment toxicity, sediment characteristics, benthic species composition, fish community structure, contaminant levels in fish and shellfish, and external pathological condition of fish (for details, see THE working group white paper at http://science.nature.nps.gov/im/monitor/COREparamMarine.doc:

## Use Neutral Criteria to Select Other Vital Signs

The monitoring plans should detail other parameters selected for monitoring and why they were selected. What neutral criteria were used in the selection process?

In considering the value of parameters as “vital sign indicators” the networks should consider criteria suggested in:

1) Park Service I&M guidance (http://science.nature.nps.gov/im/monitor/epacrit.htm#phase1 and http://science.nature.nps.gov/im/monitor/#VitalSigns and http://science.nature.nps.gov/im/monitor/monplan.doc.
2) EPA’s more detailed guidance on Ecological indicators (Jackson, L. E., J. C. Kurtz, and W. S. Fisher, editors. 2000. Evaluation Guidelines for Ecological Indicators. EPA/620/R-99/005. Research Triangle Park, NC; http://www.epa.gov/emap/html/pubs/docs/resdocs/ecol_ind.pdf and http://science.nature.nps.gov/im/monitor/phase3) and biological indicators of watershed health (http://www.epa.gov/bioindicators/).
3) The Natural Research Council (National Research Council. 2000. Ecological Indicators for the Nation” National Academy Press, (http://www.nap.edu/books/0309068452/html/, see especially chapter 4).

**Once workshops and other brainstorming and polling techniques have identified a large list of potential vital signs to measure, it is suggested that the network convene a smaller group of unbiased experts to make an initial assessment of advisable vital signs using simple neutral criteria such as those recommended above. Researchers or monitoring groups that have a pet project should not be included in this smaller group. In other words, the group should be made of those that don't have "a dog in the fight."**

**Although planners will typically end up throwing out many potential vital signs identified in workshop brainstorming, records of such decisions should be kept to show congress and others what the NPS really needed to do to answer important questions, but could not be done because of currently modest funding levels. This is particularly important when vital signs otherwise considered critical to the overall goals of the program were thrown out strictly due to cost considerations.**

**Different sets of criteria will be optimal for different networks. Some of the more common criteria that might be used are discussed briefly below. For more details, see Appendix III-A.**

**The detailed study plan should show evidence that the following concepts were considered in the selection of Vital Signs and associated measures:**

**Use Neutral Criteria in Selecting Parameters to Measure**
**Select Parameters Useful in Answering Questions**
**Select Parameters Relevant to Values to be Protected**
**Select Parameters That are Logical Parts of Multiple Lines of Evidence**
**Select Direct Measures of Specific Causes of Impairment**
**Consider Parameters Commonly Measured By Other Groups**
**Select Measures with Known and Moderate Variance**
**Select Practical and Measurable Parameters**
**Select Simple and Explainable Parameters**
**Select Relevant Forms of Parameters**
**Select Parameters Useful in Observed to Expected (O/E) Ratios**
**Consider Composite Samples to Minimize Cost and Integrate Variability**
**Select Parameters Having Regional Data Sets Collected and Analyzed the Same Way (Using Identical Protocols to Ensure Data Comparability)**
**Consider Integrative Biological Response Variables**

**Each of these is introduced briefly below. For more detail and helpful references see Appendix III-A:**

## Select Parameters Relevant to Values to be Protected

**The parameters monitored should be relevant to previously identified values to be protected. First estimate the (model) uncertainty (at least as low, medium, or high) in the relationship between the proposed vital sign and the priority values to be protected. The San Francisco Bay Network of the National Park Service placed emphasis on this concept, stating "Since there needs to be a strong scientific and**

**ecological basis for monitoring a specific indicator, the ecological significance criteria will carry the highest weight" (for more details, see appendix III-A).**

## Select Parameters Useful in Answering Questions

**Vital Signs and other parameters to be measured should be those that provide information helpful in answering the detailed questions identified in step II-C. In some regulatory applications, what will be measured, how many samples are required, minimum sampling frequencies, and (sometimes) even guidance on choosing locations, can depend on regulatory requirements. Check with the applicable State or federal regulatory agency before proceeding.**

**Consider excluding those variables that would answer no priority questions identified by the network. As explained by EPA, "The indicator must provide information that is relevant to societal concerns about ecological condition (conceptual relevance, for more detail see discussion at http://science.nature.nps.gov/im/monitor/epacrit.htm).**

## Select Parameters Having a Low Measurement Uncertainty

**Consider excluding potential vital signs and other measures with a excessively high expanded uncertainty, say those over plus or minus 30%, or whatever alternative threshold is picked. First calculate NIST expanded measurement uncertainty (not that hard to do, see step IV-F.2 for details) for each potential vital sign, and then throw out those with measurement uncertainty so high that changes or trends at appropriate effect sizes to detect thresholds of concern could never be detected, or that would require so many samples that costs would be prohibitive. See Appendix III-A and step IV-F.3 for more details.**

## Select Parameters that are Logical and Convincing Parts of Multiple Lines of Evidence

**When practicable, more than one line of evidence should be assembled to help answer questions such as whether or not habitats are impaired. Many states require such multiple lines of evidence (EPA. 2002. Consolidated Assessment and Listing Methodology, Toward a Compendium of Best Practices, First Edition, http://www.epa.gov/owow/monitoring/calm.html). Whenever possible, state or regionally recognized multiple lines of evidence should be used.**

**Regardless of whether or not they are endorsed or required by a State, multiple lines of evidence are more widely accepted as convincing when one is assessing impairment in field environments. Unlike the lab, one cannot control all the variables except one in the field, so multiple lines of evidence are typically required to get hints at causation. The Park Service typically leaves impairment decisions up to individual parks, but emphasizes the use of multiple lines of evidence in making those decisions (http://www2.nrintra.nps.gov/ard/docs/nrimpairment.pdf).**

**See Appendix III-A and EPA guidance (EPA. 2000. Stressor Identification Guidance, EPA-822-B-00-025, available at http://www.epa.gov/ost/biocriteria/stressors/stressorid.pdf for more details on the use of multiple lines of evidence.**

## Consider Specific Causes of Impairment

**If impairment is driving regulatory-related monitoring, the specific reason(s) for impairment in the area need to be considered in deciding what to monitor. Depending on the type of habitat, impairment is most often caused by the following stressors:**

1. **Excess Sediment/Siltation**
2. **Nutrients (includes TN, TP, Chlorophyll a, Nitrates, etc.)**
3. **Metals (most often mercury) in fish tissues**
4. **Dissolved oxygen (usually linked to enrichment)**
5. **Habitat and hydrologic alteration**
6. **Pathogens**
7. **Organic Enrichment/Low Dissolved Oxygen/Oxygen Depleting Organic Wastes,**
8. **Thermal Modifications (mostly from cooling towers),**
9. **PCBs in fish tissues**
10. **Priority toxic organic chemicals (like DDT and Chlordane**
11. **Fish consumption advisories in general,**
12. **Metals in Water or Sediment**
13. **Biological Criteria Exceedances**
14. **Flow Alternations**
15. **Noxious Aquatic Plants**
16. **Ammonia**
17. **Salinity/Chlorides/TDS**

**In Minnesota and several other States, mercury in fish tissues is by far the most common cause for impairment. Due to worldwide and regional air sources of mercury, enough mercury falls out of the air to cause a problem in most places in the U.S., but mercury in fish tissues tends to be extremely elevated only in areas where pH, dissolved organic carbon, oxidation-reduction (REDOX), and sulfate concentrations are optimal for the production of methyl mercury by sulfate reducing bacteria.**

**When considering impairment, monitoring planners should consider not only what States consider officially impaired according to the mandates of the Clean Water Act, but also more park-specific concepts of impairment used in the Park Service (http://www2.nrintra.nps.gov/ard/docs/nrimpairment.pdf):**

> **NPS Management Policies 2001 leave determinations of impairment to the responsible park manager and only direct that an action should be considered to constitute impairment if, in the manager's professional judgment, the action**

**"would harm the integrity of the park resources or values, including the opportunities that otherwise would be present for the enjoyment of those resources or values." NPS policies further state that whether an impact meets this definition (i.e., would harm the integrity of the park resources or values) depends on multiple lines of evidence such as**

1) **the particular resources and values that would be affected;**
2) **the severity, duration, and timing of the impact;**
3) **the direct and indirect effects of the impact; and,**
4) **the cumulative effects of the impact in question along with other impacts that are in existence.**

**See Appendix III-A for additional detail on the most frequent reasons for impairment in different habitats (rivers, lakes, estuaries, etc.).**

## Consider Parameters Commonly Measured by Other Groups:

**In addition to the "required" parameters (summarized above), what other parameters are typically important? In other words, what other things should water-monitoring networks routinely consider?**

**Depending on the particular questions, issues, priorities, and the needs of the network, other typically important (but optional) basic parameters that should probably be routinely considered by networks might include:**

**Location, date, time, and depth of collection**

**Flow or Water Level (quantitative best, qualitative better than nothing)**

.

**Habitat Measures and Observations**

**Nutrients including total nitrogen (TN), total phosphorus (TP), nitrates, ammonia, orthophosphate**

**Fecal Bacteria Measures such as fecal coliforms, E. coli, or some other alternative (Enterococci bacteria), depending on what is commonly monitored by the State. On July 21, 2003, the U.S. Environmental Protection Agency (EPA) published a rulemaking in the federal register that promulgated EPA's approval of test methods for the analysis of Escherichia coli (E. coli), Enterococci, Cryptosporidium and Giardia in fresh ambient water matrices. In addition, EPA approved test methods for the analysis of Enterococci in marine ambient water matrices (for details, see http://www.epa.gov/fedrgstr/EPA-WATER/2003/July/Day-21/).**

**Water hardness**

**Alkalinity and/or ANC.**

**Turbidity:**

> **Note: Different turbidity meters or probes can give very different results. This can limit reproducibility precision (between different meters). This brings up the importance of calculating measurement uncertainty taking into account systematic error and also reproducibility between meters. If measurement uncertainty is not calculated, one should at least account for uncertainty with simple rounding rules. This is second best, but better than nothing. For example, if more than one meter is used, round only to the number (plus one) of significant figures at which the different meters give results that round to the same values in repeat precision measurements of the same sample. Greater than two to three significant figures is seldom justified (for details see rounding rules discussion in Section VI-I. Uncertainty in Accuracy). Additional detail on issues relating to turbidity may be found in appendix III-A.**

**Secchi Disk and other water "clarity" measurements**

**Toxics such as PCBs and mercury in sediments and biota.**

**Co-factors and normalization parameters for interpretation of toxics data, such as such sulfate, total organic carbon, sediment grain size, and sediment acid volatile sulfides.**

**Photosynthetically Active Radiation (PAR):**

> **The marine work group initially recommended that light attenuation as measured by photosynthetically active radiation (PAR) be a required parameter. The group still strongly recommends that serious consideration be given to monitoring this parameter, especially in shallow environments where light penetration to the bottom is an important issue. However, there was more debate on PAR compared to the other candidate required parameters (for more detail, see the white paper at http://science.nature.nps.gov/im/monitor/COREparamMarine.doc).**
>
> **During the document review process, it was eventually decided to make PAR strongly recommended but not required at every site, due to**
>
> 1. **The expense of the instrument, and**
> 2. **Difficulties such as those outlined above, and**

3. **A less universal need to measure PAR in certain deep water oceanic environments.**

**When PAR is not to be measured, it is strongly recommended that either Secchi disk depth or turbidity be measured. The US EPA's National Coastal Assessment is implementing, in cooperation with coastal states, a more comprehensive suite of core indicators that the NPS may want to consider as additional funds become available or implement within park waters in conjunction with the appropriate state agency that is involved with the National Coastal Assessment (See white paper reference give above, op cit.).**

**The above-listed parameters are considered to be important by many other agencies (NRC, USGS, and EPA). They are also often related to or the cause of water quality impairments in NPS waters Therefore, if a detailed study plan showed no evidence that they were at least considered. If the study plan shows no such evidence WRD reviewers may request a rationale justifying why these parameters were not considered.**

**See appendix III-A discussion entitled "Consider Parameters Commonly Measured by Other Groups" for more detailed discussions of each commonly measured aquatic parameter.**

## Throw Out Monitoring Being Adequately Covered By Others in the Area

**Although it is helpful to consider parameters others consider important, it may also be important not to exactly duplicate efforts in the same area, since funding is limited. Thus, networks should consider excluding variables and/or monitoring sites where someone else (such as USGS, NOAA, FWS, or States) are already providing the data of interest or can be induced to do so in money and time-saving cooperative/partnership efforts with the NPS.**

## Select Variables Responsive to Stressors

**The plan should document the utility of the proposed measurement variables and metrics to be measured in differentiating between sites impaired and unimpaired by various stressors of concern. Consider excluding variables that have relatively little discriminatory ability. For example EPA suggests that when picking indicators, one should consider, "The ability of the indicator to discriminate differences among sites along a known condition gradient should be critically examined" (http://science.nature.nps.gov/im/monitor/epacrit.htm#phase3).**

**For more information, see EPA guidance at on stressor identification (EPA. 2000. Stressor Identification Guidance, EPA-822-B-00-025, available at http://www.epa.gov/ost/biocriteria/stressors/stressorid.pdf) and biocriteria, available at http://www.epa.gov/owow/monitoring/rbp/).**

## Select Measures with Known and Moderate Variability

**When possible, vital signs or other measurement variables picked should be those with known temporal and spatial variability characteristics, so that changes beyond normal (unimpaired condition) changes can be more easily detected without an excessive (expensive) number of samples. Generally speaking, preferred variables to be monitored and metrics to be reported are those characterized by**

**Low and/or consistent variability at unimpaired sites and/or**

**Strong changes from normal baseline variability at sites impacted by stressors of concern.**

**EPA suggests a "response variability" criterion: "It is essential to understand the components of variability in indicator results to distinguish extraneous factors from a true environmental signal" (http://science.nature.nps.gov/im/monitor/epacrit.htm#phase3).**

**Although variability characteristics are important, keep in mind that variability is not the only (or even the most important) factor in choosing variables. Some metrics with low variability may not be responsive to impacts or central to the questions at hand. If the variable in question is the most important one or the reason for the study in the first place, and it shows a high degree of variability, one might have to greatly increase sample size rather than just throwing out the variable.**

**Consider throwing out those proposed vital signs with very high true variability (sample heterogeneity) in pristine habitats. Extremely high variability may indirectly result in the need for so many samples that monitoring costs would be prohibitive. Consider sampling only in strata with lower and/or fairly homogeneous (uniform) variability. For example, if you are interested in trends and if the variability of metals concentrations in back-water pools is very high, consider sampling sediments in low gradient riffles, where variability is lower.**

**One way to assess variability within a proposed sample period is to require that a subset of sites (10 to 15 percent) be revisited within a single sample period and/or across years, as is suggested by the EMAP large river protocol (EPA 2000. Environmental Monitoring and Assessment Program-Surface Waters: Field Operations and Methods for Measuring the Ecological Condition of Non-Wadeable Rivers and Streams (http://www.epa.gov/emap/html/pubs/docs/groupdocs/surfwatr/field/R5_remap.pdf) . The re-sampling might reveal that the variability characteristics within the sample period is not very consistent.**

**The thoughts summarized above may seem obvious, but a study of flawed/failed monitoring projects revealed that many problems could have been avoided if the project designers had had a clear understanding of the temporal and spatial variability and "magnitudes of response" of parameters to be measured**

**(L.M. Reid. 2001, The epidemiology of monitoring. Jour. Amer. Water Resources Assn. 37(4): 815-819).**

**For more information, see appendix III-A.**

## Select Practical and Measurable Parameters

**It is also helpful to keep in mind that ambient conditions measured should be within the performance range of the measurement protocol selected and that methods selected should be those that can low or moderate measurement uncertainty (imprecision and/or systematic error/bias are not too great, see section VI-I).**

**Ambient monitoring, by design, tends to focus on the reduction of error associated with specific methodology so that related changes to the environment due to a pollution source will not be masked by method-quality variability (J. Diamond et al. 2001. Towards a definition of performance-based laboratory methods. A position paper of the National Water Quality Monitoring Council Methods and Comparability Board, Technical Report 01-02, Web: http://water.usgs.gov/wicp/acwi/monitoring/nwqmc).**

**Consider throwing out those variables or sites that would cause the budget to be exceeded or are otherwise unfeasible.**

**EPA suggests a feasibility of implementation criterion as follows "Adapting an indicator for use in a large or long-term monitoring program must be feasible and practical. Methods, logistics, cost, and other issues of implementation should be evaluated before routine data collection begins" (http://science.nature.nps.gov/im/monitor/epacrit.htm#phase2).**

1.

## Select Simple and Explainable Parameters

**When possible, the things being measured or counted should be kept simple (the KISS approach), explainable, and understandable. An attempt should be made to monitor things that matter to and are easily explainable to Park Service managers and other regional decision-makers. Consider throwing out those variables that are so esoteric that one could not explain their importance (and practical management changes that should be taken if critical thresholds are crossed) to superintendents and laymen.**

**EPA suggests using an "interpretation and utility" criterion, stating "A useful ecological indicator must produce results that are clearly understood and accepted by scientists, policy makers, and the public" (http://science.nature.nps.gov/im/monitor/epacrit.htm#phase3).**

**Those who have studied the success of long term monitoring programs have discovered those that many that have been most successful, produced the data most often used, and have survived long term budget cutting cycles, have tended to produce simple data that is easily understood by various user groups (Lyman McDonald, West Inc., Personal Communication, 2002).**

**However, keep in mind that although trying to use simple observations or metrics is a good idea, sometimes the most important and biologically relevant**

**parameters need to be measured, even though they are not particularly simple. Keep track of those important things that couldn't be monitored for cost reasons, for explanations to congress.**

## Select Relevant Forms of Parameters

**Relevant forms of parameters should be measured. To compare with a given standard, one may need to specify wet weight rather than dry weight, or one may need to specify total unionized ammonia rather than total ammonia. Any potential inability to determine the relevant forms of the parameter being measured should be considered during project planning and the plan should detail the rationale for relevancy (to values being protected and desired future conditions) of the parameters being measured.**

**Reporting units (for example ug/L in water or ug/kg both wet and dry weight in tissues and sediments) and normalization parameters (for example, sediment samples normalized to acid volatile sulfides or total organic carbon, or flow-normalized concentrations) should be specified in the plan. It is often best to report flow-adjusted concentrations when trends in moving water are to be considered and concentrations are found to correlate with flow (See step III-A).**

**Concentrations of certain contaminants are sometimes normalized by percentage of fat content of the tissue.**

## Select Parameters Useful in Observed to Expected (O/E) Ratios

**The ratio of "observed to expected" (0/E) number of taxa present is one example of a metric that can be practical and easy to explain. Another advantage is that O/E metrics is that such metrics can be rolled up into nationwide condition indices easier than some other parameters. This facilitates understanding of mangers of the meaning of changes.**

**If vital signs monitoring results are to be expressed on O/E rations, the vital signs chosen and the study designs need to be coupled in a methodical way to allow 0/E comparisons.**

**The National Research Council recommended using total species richness (in terms of observed to expected ratio (O/E), or what number of taxa are present relative to the number that would be expected if human impacts were not present. Other ecological indicators recommended by NRC included native species richness (number of taxa), land cover and land use, nutrient runoff, soil organic matter, primary productivity (chlorophyll a, etc.), lake trophic status, stream oxygen, and nutrient use efficiency. All can utilize O/E ratios (National Research Council. 2000. Ecological Indicators for the Nation" National Academy Press, (http://www.nap.edu/books/0309068452/html/, chapter 4).**

**O/E ratios can be very simple. For example, if seining produces at least 14 species of small fish at most sites along a river, but only 3 species of fish can be collected that way at a similar but effluent-dominated site just below a sewage treatment plant, the O/E ratio for small (seine-collected) fish at that impacted site**

**would be 3/14, reflecting reduced biodiversity and the fact that only the really tough species can survive the effluent effects. Regulators understand simple metrics like that, and sometimes that is all it takes to convince them they need to improve the water quality.**

## Consider Composite Samples to Minimize Cost and Integrate Variability

**Composite samples should not be used for volatile chemicals and should be used with great caution, if at all, for lighter semi-volatile chemicals such as naphthalene. In other cases, composite samples can often help integrate values in different locations and thus reduce the number of samples needed and/or the variability. They should be used only when representativeness is not compromised. If they are to be used, the plan should present a rationale justifying why the pros outweigh the cons (such as loss of temporal or spatial specificity and potential mixing of separate target populations) in the particular situation should be summarized in the plan.**

**Sometimes one will have to composite non-volatile samples to achieve data comparability. For example, USGS water column collections are often composited by depth and cross section. As emphasized in section VI-A, data comparability is desirable and when possible, the NPS should cooperate with other agencies in collecting comparable data, so that data collected is coordinated and consistent and so that mangers can use combined datasets to answer questions and make management decisions.**

## Select Parameters Collected and Analyzed the Same Way By Others in the Region to Ensure Data Comparability

**As discussed in more detail in section VI-A, if chemical concentrations are to be measured for State or federal regulatory purposes, the team should make sure the methods used produce comparable data and are acceptable to the appropriate regulatory agencies.**

**Consider throwing out variables for which regional protocols are not already available and thus data comparability may be a factor. Data comparability is a quality assurance basic, and if there are no comparable data sets, one is often left wondering if some regional effect (cold year, hot year, dry year, wet year, regional air pollution, sun spots) is a key stressor rather than a local stressor.**

**Other things being equal, throw out variables which require protocol development or are being addressed to answer research questions rather than monitoring questions. For aquatic monitoring, many standard monitoring protocols are already developed and are widely used, so developing totally new protocols can seldom be justified.**

## Consider Integrative Biological Response Variables

**To answer vital signs general (long term) monitoring questions about what is changing in response to known or to unanticipated stresses, integrative biological response variables, especially those specified by the States for Biocriteria monitoring, should be measured to the extent practicable.**

**Integrative biological response variables include the status of aquatic invertebrate populations, the status of fish populations, and chlorophyll a. These are examples of variables that respond to multiple stresses (from excess nutrients, toxics, and other stresses) and also reflect integrated (combined) effects over time. Biological response variables are also often quite relevant to the environmental value is be protected/desired future condition (discussed in step II-B).**

**For additional detail, see appendix III-A.**

## III-B. Identification of Decisions and Decision Rules

**What are the potential decisions that hinge upon the monitoring results and what types of results would trigger various identified decisions? Typically, pre-project decision rules should be stated as an "If-then" decision rule. The "then" part would be actions the park would take if certain results were obtained.**

**An example decision rule is provided as follows:**

> **Nesting birds in the area may be impacted by human disturbance from a new trail. If reproductive success in this area is reduced by 20%, we will then close the trail and keep monitoring to determine if our management action has helped improve reproductive success.**

**Depending on the issue, the action might be working with a regulatory agency to effect a change (for example, through the TMDL process, or something as simple as deciding to continue long-term monitoring.**

**Measurement sensitivity (such as detection levels, see step V-B.4) needs to be fine enough in scale, and sample size/frequency (relates to power, steps III-B and V-C) needs to be adequate to make it possible to detect changes considered to be ecologically significant, or changes large enough to be beyond "bio-equivalent" ranges. See also, discussion on bioequivalence testing, section IV-C).**

**In NEPA settings, Park Service managers are asked to decide what an impairment would look like, deciding, in the managers' professional judgment, whether or not a NEPA action would "harm the integrity of Park resources for values, including the opportunities that would otherwise be present for the enjoyment of those resources or values " (http://www2.nrintra.nps.gov/ard/docs/nrimpairment.pdf).**

**In the more general sense, it is likewise valuable to decide how big of a change (effect size) would be big enough to be considered too big and therefore not tolerable. Again, in the ecological sense, this often comes down to changes considered to be ecologically significant, or changes large enough to be beyond "bio-equivalent" ranges.**

**A 40% effect on the larvae of a ubiquitous species of mosquito may be far less problematical than a 5% effect on a long-lived endangered mammal, partly**

**because the mosquito produces more young and the young produced are subject to high mortality even in pristine areas. Endangered species are highly valued by society and laws and are therefore protected at the individual level, whereas most species are protected at the population level.**

**When possible, the plan should define differences that are considered biologically meaningful vs. differences within bioequivalent ranges. If such levels are not known, the monitoring should be designed in such a way that threshold levels of change considered "too much" are developed.**

**Trigger points or threshold values are typically related to values considered to be beyond "de minimis", an abbreviated form of the latin phrase "de minimis non curat lex" which translates to "the law cares not for small matters." In the risk assessment or environmental assessment, the phrase de minimis has been used in the context of a risk or effect-size which is small enough to be negligible (see Appendix III-B for more detailed discussion.**

**A 1992 paper suggested that it was difficult to find cases where a state or federal regulatory agency had prosecuted anyone for a biological effect size of less than 20% on non-human or non-endangered species. This was true regardless of whether the population, community, or ecosystem level was being considered (Suter, G.W. II, A. Redfearn, R.K. White** and **R.A. Shaw. 1992. Approach and strategy for performing ecological risk assessments for the Department of Energy Oak Ridge Field Office Environmental Restoration Program. Martin Marietta Environmental Restoration Program Publication ES/ER/TM-33, Environmental Restoration Division Document Management Center Environmental Report (ER), Environmental Sciences Division (ESD) Publication 3906, Oak Ridge National Laboratory, Oak Ridge, TN, pp. 8-9**

**However, like endangered species, rare resources in National Parks are rare and special and perhaps deserve "special" protection. The Park Service is not necessarily accepting 20% hits on these resources any more than the FWS is accepting 20% hits on endangered species (for more detailed discussion of de minimis concepts, see appendix III-B.**

**"Special protection" could be protection at the individual level (we don't even want to lose one) rather than the population level, and could involve more stringent than normal criteria or state standards for levels of contaminants in water, prey tissue, sediments, or soils. Special protection for endangered species and highly valued resources in National Parks is recognized in risk assessment (EPA 1998: Guidelines for Ecological Risk Assessment Risk Assessment Forum, U.S. Environmental Protection Agency Washington, DC, EPA/630/R-95/002F, April 1998, Final (www.epa.gov/ncea/ecorsk.htm).**

**Often decision rules will be triggered by concentrations (trigger points or thresholds) of contaminants that exceed applicable water quality standards or other comparison benchmarks, such as those found in the sources summarized below:**

## Sources of Data Comparison Benchmarks Useful as Trigger Points or Thresholds:

**Water quality standards are published by each State, typically updated each 3 years, and these standards are often available on State Websites. Many State Water Quality Standards are also available as links from EPA's Water Quality Standards Database (http://www.epa.gov/wqsdatabase/overview_inter.html). Simply consult the Enviromapper map of states with standards available on the internet and then click on those state of interest at http://www.epa.gov/waterscience/standards/wqslibrary/states.html.**

**More generic national water quality criteria (upon which many States base water quality standards) for various water uses and various contaminants are published by EPA (http://www.epa.gov/waterscience/standards/about/crit.htm). If water is used for drinking water, maximum contaminant level (MCL) criteria for drinking water are published by EPA (http://www.epa.gov/safewater/mcl.html).**

**Marine sediment, freshwater sediment, and soil benchmarks are available from NOAA in handy Screening Quick Reference Tables (SQuiRT) format http://response.restoration.noaa.gov/cpr/sediment/squirt/squirt.pdf). More recent references not in the SQuiRT tables and evidently not on the Internet are (Chris Ingersoll, Columbia Lab, USGS, Personal Communication, 2003):**

**Smith SL, MacDonald DD, Kennleyside KA, Ingersoll CG, Field J. 1996. A preliminary evaluation of sediment quality assessment values for freshwater ecosystems. *J Great Lakes Res* 22:624-638.**

**MacDonald DD, Ingersoll CG, Berger T. 2000. Development and evaluation of consensus-based sediment quality guidelines for freshwater ecosystems. *Arch Environ Contam Toxicol* 39:20-31.**

**Oak Ridge Ecological Risk Assessment Screening Benchmarks are available on the internet at http://risk.lsd.ornl.gov/homepage/eco_tool.shtml.**

**Recent sources of toxicity profile information include Oak Ridge (http://risk.lsd.ornl.gov/tox/rap_toxp.shtml) and (for human health) ATSDR (http://www.atsdr.cdc.gov/toxpro2.html). Ron Eisler's Contaminant Hazard Reviews (with emphasis on effects on fish and wildlife, summarized for 35 contaminants) are now available on the Internet (http://www.pwrc.usgs.gov/cgi-bin/om_isapi.dll?clientID=842570&infobase=eisler1.nfo&softpage=Browse_Frame_Pg).**

**Summaries on data comparison benchmarks for metals and industrial organics and petroleum hydrocarbons in water, sediment, soil, and tissues, updated through 1997-1998, are summarized in the NPS contaminants Encyclopedia (www.nature.nps.gov/toxic). Many of the documents listed above that were available in 1997 were quoted. Although the NPS encyclopedia has not been updated since 1998, it contains information not available in the other documents. The NPS encyclopedia also contains general ecological toxicity profile information on 118 contaminants.**

## Typical NPS Quality Assurance Objectives:

**Key decisions that that depend on the answers to study questions shall be identified in the plan.**

**The plan shall include an identification of threshold results (effect sizes) that would trigger identified conclusions, decisions, or actions. When practicable, pre-project decision rules should be stated as an "If-then" decision rule. The "then" part would be actions the park would take if certain results were obtained.**

**When biological thresholds are being considered, inequivalence procedures should be used when practicable (see appendices for more detail).**

**Part of decision rules should relate to the maximum amount of uncertainty that would be tolerable. Obviously a plus 5% effect we be difficult in not impossible to document with a reasonable sample size and sampling frequency if total uncertainty from measurement precision and measurement systematic error (bias) factors add up to a plus or minus 40%. Initial decisions related to maximum allowable uncertainty should be made at this step, and revisited in more detail in later measurement and model uncertainty decisions (see section IV-F.2 and IV-F.3).**

**Accordingly, when initial (pilot scale) monitoring results indicate an excessive lack of measurement precision (excessive imprecision) or an excessive amount of measurement systematic error (bias), or other general signs of a general lack of accurate measurability, or when initial results indicate a lack of ability to detect trends in the face of excess daily or seasonal true variability vs. budget-limited infrequent monitoring, the monitoring plan should be changed to make sure that future monitoring will be effective in helping to answer study plan questions. Any pilot study monitoring results that would trigger such decisions should be identified in the plan.**

**The plan should document the consequences of wrong-decision errors and use this analysis to set initial goal limits on decision errors. When a hypothesis testing approach is used, EPA recommends that the data quality objective process step of setting limits on type I and type II decision errors be initiated right after considering the decisions that depend on the data. However, our long term monitoring in the NPS is not limited to hypothesis testing designs, so our main emphasis on bounding error and uncertainty comes towards the end of the planning process (Step IV-F.2). Nevertheless, it is a good idea to begin to make some initial estimates of tolerable error rates at this early stage. The consequences of making a wrong decision as a type II error (concluding there is no impact when there is) is often more serious for an endangered species or a rare resource in a National Park than for less rare or valuable resources. Therefore type II error rates (beta) are often set to no more than 5% when possible (1% is better and 10% the maximum). See step IV-F.2 for more detailed explanations.**

**Chemical concentrations (in water, sediments or tissues in the identified target population) that would trigger significant concern, and the specific regulatory or other decisions/actions such results would trigger, should be identified in the plan.**

**For additional details, see section on bioequivalence (step IV-C, below) approaches and appendix III-B of this document as well as EPA guidance at http://www.epa.gov/quality/qs-docs/g4-final.pdf.**

## III-C. Peer Review of Phase II

**The description of this peer review step is placed here to help networks understand how this VS program-required peer review step fits sequentially into the other general and QA/QC monitoring planning steps.**

**A second round of peer review and approval of the Phase 2 report should occur after the initial list of vital signs and measurable objectives is determined, and prior to detailed work on sampling design, protocol development, database design, etc. The phase 2 report will include will cover (the rest of) Step 3 as well as Step 4 of the 7-step recommended approach. Step 4 is "Write a report on the workshop and have it widely reviewed." The Phase 2 report should be a draft of Chapters II, III and IV of the monitoring plan.**

**The Board of Directors will typically not have the technical expertise to develop the details required for good statistical design and adequate quality assurance. Therefore, the small-group project planning team, ordinarily consisting of the Science Advisory Committee plus technical experts with more detailed and specialized expertise in practical statistics, study designs, and QA/QC, should again convene for more detailed planning. After receiving the general advice of the Board of Directors, the smaller group should again gather to discuss and document decisions related to the following basic QA/QC planning steps.**

**The results of this peer review should be helpful to a network that is attempting to fine tune initial decisions on sampling design, protocol development, database design, etc. After the phase 2 report is drafted and peer-reviewed, modifications in the report shall be made based on peer review, before proceeding to phase 3.**

## PHASE 3 FOR THE DEVELOPMENT OF VITAL SIGNS NEWORK MONITORING PLANS

**Now that the first four steps of NPS VS general (not just aquatic) guidance (http://science.nature.nps.gov/im/monitor/approach.htm), have been completed, the small group of technical experts (usually the Science Advisory Committee plus technical experts with more detailed and specialized expertise on practical statistics, study designs, and QA/QC) should again gather to plan phase 3. A goal of phase 3 will include optimizing the study design once the steps above and been completed, developing the proposed final monitoring plan. Phase 3 tasks include steps 5 to 7 of the basic VS strategy (see phase III discussion at end of appendix IV and VS guidance at http://science.nature.nps.gov/im/monitor/approach.htm):**

**Hold one or more meetings to decide on priorities and implementation approaches.**

**Draft the monitoring strategy.**

**Have the monitoring strategy reviewed and approved.**

# IV. SAMPLING DESIGN

## IV-A. Overall Sampling Design

**The plan should detail how the basic design is optimal to answer identified questions considering identified target populations, study boundaries, and sample units. The overall statistical design needs to be carefully thought through and documented in the plan before monitoring begins.**

**Typically one cannot afford to sample everything at all potential sites and all potential times, so one typically samples a limited amount of times and locations and then tries to make statistical inferences (conclusions) about the larger target population.**

**If monitoring is being done partly to get hints about causation, or even correlations between environmental variables, suggestions in EPA's stressor identification guidance is helpful. In monitoring being done for these purposes, planners need to carefully consider optimal frequency and location of monitoring, and which suspect co-factors to monitor. EPA defines a cause as "a stressor that occurs at an intensity, duration, and FREQUNCY of exposure that results in a change in the ecological condition." EPA also points out that "Often associations between candidate causes and effects can be improved by identifying and isolating confounding factors in either the receptors or the environment. For example, the frequency of hepatic neoplasms in fish is associated both with the age structure of the fish population and the concentration of PAHs in sediment…Similarly, a decline in fish species richness is a common measure of impairment, but the number of species present generally increases with increasing stream size. Therefore, including a correction for stream size could strengthen the association between the degradation and species loss…Measures of exposure from the case at hand can also be matched with measures of effect from other situations. The objective of this analysis is to provide evidence showing that the stressor is present at the study site in sufficient quantity or frequency that the investigator would expect to see a particular effect based on effect information from laboratory tests, field tests, or exposure-response relationships developed at other sites…The constancy of association needs to be considered. For example, is the constancy invariant at many places and times and at background frequencies are there many exceptions to the association? If values exceed a regulatory criterion or threshold value, do they also exceed that value at pristine sites with a similar frequency?" (http://www.epa.gov/ost/biocriteria/stressors/stressorid.pdf).**

**Typical NPS quality assurance and data quality objectives:**

**(http://science.nature.nps.gov/im/monitor/monplan.doc).**

**The monitoring plan should document how the selected overall study design is optimal for providing information adequate to answer identified questions, considering the target population and temporal and spatial boundaries of the population to be monitored.**

**Sampling design details that need to be documented in the plan include where, when, how large a sample size, and how often selected vital signs will be measured (frequency).**

**In concert with general VS guidance, the plan shall therefore include an explanation of how the overall statistical sampling design will allow for inferences to be made to areas larger than those actually sampled. The plan shall identify populations to be sampled and sampling units. For each park, a description should be provided of the approach used to determine where sampling will occur for each vital sign, including justification for collocating or not collocating sampling for various vital signs. The plan shall include justification for the factors used to stratify the park into sample units (e.g., cost of access, terrain features such as elevation and slope, soils or vegetation map)… The plan shall include detailed maps and descriptions of where samples will be taken should be included in the protocols or an appendix, but the plan should summarize the overall spatial monitoring design for each park (http://science.nature.nps.gov/im/monitor/monplan.doc).**

**For projects collecting considerable new geospatial data (derived from remote-sensing, mapping, and surveying technologies) the plan shall detail how QA/QC for geospatial work has been implemented (EPA 2003, Guidance for Geospatial Data Quality Assurance Project Plans (QA/G-5G), EPA/240/R-03/003 http://www.epa.gov/quality/qs-docs/g5g-final.pdf).**

**Typically the plan should describe what is known about average values and variability in the various strata.**

**The plan shall detail and how the sampling scheme will insure that the value obtained will be representative of the target population being studied. Representativeness is important to answering questions and basic study design, and is so important that it should be considered both here and in more detail in step V-D on representativeness.**

**If the variability and typical values in various potential strata is not well understood, the plan should identify how pilot scale monitoring or literature review efforts will be initiated to determine these values before the monitoring design is finalized.**

**Timing and frequency of sampling needs to considered in relationship to how variables are known to vary over time. In some cases, low flow periods or summer periods of sampling can be justified as worst case or relatively stable periods for detecting changes over the years. However, unless one understands how the variability (typically as a standard deviation) of parameters in pristine or un-impacted areas changes according to flow, season or even time of day, it is difficult to design a monitoring frequency optimal for detecting trends or correlations with other factors such as land use.**

**If the monitoring is to be done to generate general status and trends information or to compare conditions at sites known or suspected or being impacted versus relatively pristine regional control sites, the plan shall document how the overall study design is optimal for those purposes.**

**The team developing and reviewing the detailed study plan and QA/QC steps should include a professional statistician or a water professional who is very familiar with applied environmental statistics and applied environmental survey/monitoring design. This expert should not be just any statistician, but one with considerable expertise related to developing environmental monitoring designs in general, applied parametric and nonparametric environmental statistics in particular. If biological monitoring is to be included, the statistician should also be familiar with inequivalence testing. Before it is finalized, the plan should be approved by the team statistician(s) to make sure that the plan has been optimized to allow valid statistical analyses that will be helpful in helping answer the identified questions.**

> **Why? A study of flawed/failed monitoring projects revealed that many of the problems "could have been avoided by (pre-project) submission of the study design to thorough statistical review" (L.M. Reid. 2001, The epidemiology of monitoring. Jour. Amer. Water Resources Assn. 37(4): 815-819). Professional statisticians and study design experts can not only help networks with basic survey and monitoring design, but can also help monitoring teams avoid common junk science mistakes, like equating correlation with causation or reporting a standard error or 95% confidence interval when the sample size is three.**

## IV-B. Identification of Target Population, Study Boundaries, & Sample Units:

**The plan shall identify the target population, study boundaries in time and space, and any identified strata or other sample units. The target population is simply the larger universe of all possible values (bounded in time and space) that one is sampling from and wishes to make statistical inferences about. It often does**

**not signify population in the sense of a population as a specific level of biological organization (see appendix IV-B for more detail).**

**Stratified random sampling has been common because many cannot afford simple random sampling. A stratum is a major subdivision of the sampling universe (e. g., all first order streams) chosen to reduce the variability (Paul Geissler, USGS Patuxent Lab, Personal Communication 2002). Strata are a type of sampling unit, but one can also define smaller sample units within identified strata.**

**Sample units are typically some subdivision of the larger universe that might be sampled, even in that larger universe is itself one stratum. In other words, they are subunits of a larger area, volume, or mass of interest. When the target population is made up of "natural units," such as people, plants, or fish, then the definition of a sampling unit is straightforward. However, many environmental studies involve target populations made up of continuous media, such as air, water, or soil. In this context, the sampling unit must be defined as some volume or mass to be selected" (http://www.epa.gov/quality/qs-docs/g4-final.pdf).**

## IV-C Proposed Statistical Analyses to be Used (also relevant to section VIII, below):

**In many classical statistics textbooks, a set of observations is drawn at random, from a normally distributed population, and with constant mean and variance. Water quality data are typically not normally distributed, observations are not usually drawn at random, and the mean and variance are not constant over time. This general background behavior of water quality variables greatly influences the manner in which appropriate statistical methods are selected for analyzing water quality data (Robert Ward, CSU, Personal Communication, 2001). The absence of normal distributions is one reason that many use nonparametric methods (see section IV-C for more detail).**

**Many statistical calculators and summaries are now on the Internet. Examples of summaries include:**

1. **A power calculator and other resources are in EPA's statistical primer at http://www.epa.gov/bioindicators/primer/ and**
2. **A commonly used text book (Helsel, D.R. and R.M. Hirsch 1992. Statistical Methods in Water Resources. Studies in Environmental Science 49, Elsevier Publishing, NY, http://water.usgs.gov/pubs/twri/twri4a3/pdf/twri4a3.pdf. Other more detailed examples are listed throughout the discussion below.**

**Although some countries and states and moving towards multivariate methods, and there are often unresolved issues with multimetric methods. Nevertheless, many of the multimetric methods are still useful.**

**In considering options for statistical study designs, one needs to carefully consider related issues such as:**

1. **Statistical/study design sensitivity, pertaining to multiple data points and the ability of the statistical design to detect statistical changes considered important, is discussed in the section V-B.7.**

2. **Conceptual Model Uncertainty (see section IV-F.2), and**
3. **EPA's extensive guidance on stressor identification (EPA. 2000. Stressor Identification Guidance, EPA-822-B-00-025, available at http://www.epa.gov/ost/biocriteria/stressors/stressorid.pdf).**

**The following concepts should also be considered:**

## Consider Accepted Rules of Thumb:

**It is suggested that planners consider guidance suggested by EPA ("Using Statistics and Statistical Hypothesis Testing for Analyzing Observational Data in Stressor Identification," Chapter 3 of EPA's Stressor Identification Guidance (EPA. 2000. Stressor Identification Guidance , EPA-822-B-00-025, available at http://www.epa.gov/ost/biocriteria/stressors/stressorid.pdf). Therein:**

**EPA encourages:**

**The use of summary statistics.**

**The use of correlations or regression techniques to quantify relationships between variables**

**EPA recommends caution in using statistics:**

**To determine the probability that two sets or samples are drawn from the same distribution, or that they differ by a prescribed amount.**

**To determine the probability that a relationship is nonrandom, or that the slope of a regression differs from zero.**

**EPA advises us to avoid:**

**Using the results of statistical hypothesis tests to conclude that a candidate stressor is (or is not) the cause.**

**Depending on results when the assumptions of statistical hypothesis testing are violated.**

**Using replicate treatments in observational studies, when the replicates cannot be randomly assigned in a way that minimizes the influence of confounding variables (thereby a significant outcome in a hypothesis test may be falsely attributed to the wrong cause).**

**Concluding that statistically-significantly or correlated variables have a causal relationship.**

**Handy internet guidance documents and tools for choosing the right statistic or analysis, internet calculators, and other handy statistical links are found at websites such as the following:**

1. **http://statistics.com/content/javastat.html#WhichAnalysis**
2. **The many hypothesis testing and interval and sample size/power calculator websites listed in sections below and appendix IV-C.**

**For more details on rules of thumb and general guidance, see appendix IV-C**

## Consider Using Intervals:

**Properly framed parametric or nonparametric 95% confidence intervals, especially those adjusted for measurement uncertainty, rather than point estimates, should be used when practicable to bound uncertainties on reported means, medians and other quantiles, trend lines, or other summary statistics.**

**Consider using nonparametric intervals when the distributions are not normal. Planners need to be aware that confidence interval interpretation is subject to some of the pitfalls of hypothesis testing (see appendix IV-C for details).**

## Trend Analysis Options:

**In general, the ability to detect trends depends on three things: the strength of the signal, the variability of the data when there is no signal, and sample size. Another way to say this is "the ability to detect trends or evaluate patterns is determined by the signal to noise (sic, natural true variability) ratio of the data and sample size" [C.A. Stow et al. 1998. Long term environmental monitoring: some perspectives for lakes. Ecological Applications 8(2):269-276].**

**The difficulty of documenting trends in outdoor systems is one reason we recommend in an earlier step picking indicators that 1) respond to small changes, and 2) have low variability in natural settings, 3) have high signal to noise (sic, natural true variability) ratios. High signal to noise ratios tend to occur with indicators that have low naturally occurring variability and/or very high responsiveness to signals (Environment Canada reports at http://science.nature.nps.gov/im/monitor/emancmv.doc).**

**Since we often cannot manipulate the strength of the signal, a key reason we encourage the use of indicators with low variability in pristine (or less impacted) sites, is that low variability in natural settings typically makes it easier to pick out a change strong enough to be considered a signal rather than true natural variablity.**

**When practicable (or unless an alternative approach is justified), for general trend analyses, the traditional (for water quality work) Seasonal Kendall Test for trends, perhaps with data adjusted for measurement uncertainty and/or flow. should be considered as the default method for trend analyses. This is recommended for NPS consistency. The Seasonal Kendall Test (a seasonal extension of the nonparametric Mann-Kendall test) was found to be the most frequently used trend analysis method in water quality work, especially in USGS analyses (L.M. Griffith,**

**R.C. Ward, G.B. McBride, J.C. Loftis. 2001. Data Analysis Considerations in Producing 'Comparable' Information for Water Quality Management Purposes. National Water Quality Monitoring Council Technical Report 01-01. White Paper of the National Water Quality Monitoring Council, Co-sponsored by USGS, Web: http://water.usgs.gov/wicp/acwi/monitoring/CouncilPrior6-Mar00.html). The seasonal KT test is available as an easy option on WQStat software used broadly in USGS.**

**For more information on the seasonal Kendall Test for trends and available software, see appendix V-B.3.**

**Flow-adjusted concentrations should be considered in trend analyses IF there is a strong correlation between the concentrations and flow. See appendices III-A and V-B.3 for more detail.**

**Other normalizing factors (pH, organic carbon, grain size, acid volatile sulfides, etc.) may have to be factored in to make sure a perceived trend is really due to time rather than some other cofactor.**

## Consider Bio-Inequivalence Testing:

**In concert with the rule of thumb that Parks typically want to detect changes that are biologically or physically significant, rather than those that are merely statistically significant, inequivalence hypothesis testing designs should be used as one line of evidence, when possible. These designs should be capable of detecting a change large enough to be of concern. Those who find the statistics and/or software requirements of inequivalence testing are too challenging are encouraged to find a statistician to help.**

**In inequivalence tests, "to reject that hypothesis is to infer that the variables being sampled are very unlikely to differ by more than a specified amount (the prescribed interval width) and so may be considered as "equivalent". That is, differences *are* expected, but if they are small enough the variables can be considered as equivalent. Note that in this case the significance level (*a*) protects the consumer's risk, so it is essentially precautionary in nature"…(to be more protective). (G. McBride. 2000. Some issues in statistical inference, published on the Internet at http://www.niwa.co.nz/rc/prog/stats/intro/, click on title). See also, discussion of bioequivalence (below).**

**Otherwise, as a last resort, normal hypothesis testing can be used as one line of evidence, as long as sample sizes are large enough to ensure that statistical power is at least 95%. The National Park Service would typically want to avoid false negatives (saying there was no effect when there was) in protecting rare and highly valued resources in National Parks. Therefore, limiting beta to no more than 5% is generally a good (see section IV-C for more detail. However, using an inequivalence test is typically a more desirable approach for those capable of doing, it.**

**Equivalence and inequivalence testing relate to how data variability is to be handled. If one is trying to determine an effect caused by man or a change beyond natural variability, what change or "effect size" will be considered significant? In other words, most biological resources (and water chemistry measures) are naturally variable even in pristine settings with no influence by man. This true**

**natural variability may be considered the natural up and down variability (ture heterogeneity) in the system and is typically the desired future condition in resources that are currently impacted and thus not currently varying within natural or desirable ranges.**

**It follows then that in trying to determine an effect or change that is beyond the natural variability, it is typically necessary to determine the amount of variation that is "natural" versus what is beyond natural. Certain "de minimis" changes are so small that they do not matter. How does one determine bio-equivalent or "no significant effect levels" and thereby help differentiate background normal variablity from a separate "signal" representing a significant effect or change?**

**Inequivalence testing is more philosophically aligned with the philosophy of protecting unique and valuable resources in National Parks, as explained in various NPS statutes, regulations, and policy statements, than is a typical null hypothesis test. Inequivalence testing is also more aligned philosophically with philosophies that are compatible with protection of highly valued and rare natural resources, even in the face of incomplete scientific information, such as the "precautionary principle" and the larger universe of philosophies defined as "eco-ethics."**

**Along with ecological sustainability, the precautionary principle ("even in the absence of conclusive scientific evidence, we shall act to protect the health of humans and the environment") is one key piece of a more general statement of eco-ethics (Cairns, J. 2002. A Declaration of Eco-Ethics, Ethics In Science And Environmental Politics (ESEP):79–1, http://www.esep.de/articles/esep/2002/E20.pdf).**

**Unfortunately, the limitations in scientific methods to quantify causal relationships are often misinterpreted as proof of safety (J. A. Tickner. 2003. Precaution, Environmental Science, and the Research Agenda. SETAC GLOBE, September-October 2003, 37-38). Actually proving cause and effect is a complex process (EPA. 2000. Stressor Identification Guidance, EPA-822-B-00-025, available at http://www.epa.gov/ost/biocriteria/stressors/stressorid.pdf).**

**However, due to their precautionary nature (the burden of proof is for a proof of safety versus proof of hazard), inequivalence tests are especially useful relevant to protection of unique or rare resources in National Parks (and similar unique resources such as endangered species). In general, proof of safety is much more difficult to achieve than is proof of hazard and such proof typically demands a higher number of samples. To prove safety one is forced to reject hypothesis to a high standard of proof (typically, alpha = 5%). Inequivalence testing should thus be used when possible rather than classical null hypothesis testing.**

**As mentioned in the Decision Rules section IV-C, ecologists and mangers can raise thoughtful objections to interval testing in that it requires the *a priori* specification of the equivalence interval. Specifying the boundaries of that interval is really a matter of informed professional judgment. But a power analysis requires exactly this information also! It is simply the case that meaningful inferences require more data than classical test procedures require (G. McBride. 2000. Some issues in statistical inference, published in Internet at http://www.niwa.co.nz/rc/prog/stats/intro/ (then click on title).**

**If bioequivalence tests are to be used, a test of the inequivalence hypothesis utilizing the TOST (Two One-Sided) test procedure should be utilized. The default hypothesis should be that presumably impacted and presumably pristine sites are not equivalent. In other words, the precautionary principle should be used and the presumption should be lack of equivalence rather than the null of no difference. Those wishing to prove bioequivalence should have the burden of proof. For more information on inequivalence and equivalence testing, including computer programs that handle these procedures, see appendix IV-C.**

# Use Caution with Traditional Hypothesis Tests

**For field studies, it is important not to hint that one has proved cause and effect by showing a difference through traditional null hypothesis testing, especially when beta is not controlled. Some well known experts now say hypothesis testing should NEVER be used in field studies, because one can never control all the variables except the test variable and because the results tend to give decision-makers false assurances that something has been "proved" with a high degree of confidence.**

**The World Wide Web contains some good discussions on hypothesis testing abuses. In spite of the broadening recognition, problems with hypothesis testing have probably not been broadly (enough) recognized in the fields of water quality, general biology, monitoring, contaminant hazard assessment, or field observational studies in general. While these fields have been slow to pick up on newer thinking, hypothesis testing was being de-emphasized in some other disciplines. For more information, see Douglas H Johnson's 1999 paper (D.H. Johnson. 1999. The Insignificance of Statistical Significance Testing. Journal of Wildlife Management 63(3):763-772, found the Internet at URL http://www.npwrc.usgs.gov/resource/1999/statsig/stathyp.htm). Among other things, Johnson states that:**

> **Statistical testing of hypotheses in the wildlife field has actually increased dramatically in recent years….While this trend was occurring, statistical hypothesis testing was being deemphasized in some other disciplines. As an example, the American Psychological Association seriously debated a ban on presenting results of such tests in the Association's scientific journals…Speakers at the 1998 annual conference of The Wildlife Biometrics Working Group were virtually unanimous in their opinion that hypothesis testing was overused, misused, and often inappropriate…Statistical hypothesis testing has received an enormous amount of criticism, and for a rather long time.**
>
> **In regards to the (over emphasis) on hypothesis testing, Johnson went on to quote Clark as noting that it was "no longer a sound or fruitful basis for statistical investigation," noting that Bakan called it "essential mindlessness in the conduct of research." W. Edwards Deming was quoted as having commented that "the reason students have problems understanding**

> hypothesis tests is that they may be trying to think." Carver was quoted as having concluded that "statistical significance testing should be eliminated; it is not only useless, it is also harmful because it is interpreted to mean something else." …Loftus is quoted as having said he "found it difficult to imagine a less insightful way to translate data into conclusions."

In 2000, David Anderson summarized pervasive problems with using standard null hypothesis testing in field ecological and observational studies and proposed better alternatives (http://www.cnr.colostate.edu/~anderson/PDF_files/TESTING.pdf). See also David Anderson's other Colorado State University sites related to problems with using hypothesis testing in field and observational studies:
http://www.cnr.colostate.edu/~anderson/null.html
http://www.cnr.colostate.edu/~anderson/nester.html
http://www.cnr.colostate.edu/~anderson/thompson1.html

Never using standard null hypothesis testing in field studies or observational studies, as some recommend, may be a bit extreme. However, if hypothesis tests are used at all, perhaps as one piece of a weight of evidence presentation after other statistical models were considered but rejected, be sure to specify a reasonable degree of power (1-beta) before the project begins in addition to specifying significance (alpha) and be sure to discuss the limitations and uncertainties associated with the hypothesis testing model for the particular data set being analyzed. For especially valuable natural resources, such as natural resources in National Parks or endangered species, a power of .95 or .99 may be appropriate.

Therefore, if standard null hypothesis tests are to be done, beta should be limited to no more than 5% when possible (1% is better and 10% the maximum) and the results of the hypothesis test should be used as only one of several lines of evidence in conclusions. As mentioned above, inequivalence testing is the preferred default when hypothesis testing is to be used. Inequivalence tests with $\beta = 0.01$ will typically demand a significant sample size, though alpha can be made as high as 0.2 to ameliorate the sample size to somewhat smaller levels.

Using the precautionary principle, EPA's DQO guidance recommends using 0.01 as the starting point for setting decision error rates. If the consequences of a decision error are not severe enough to warrant this stringent decision error limit, this value may be relaxed (a larger probability may be selected) (EPA. 2000. Guidance for the data quality objectives process. EPA QA/G-4, ORD, EPA/600/R-96/055, http://www.epa.gov/quality/qs-docs/g4-final.pdf).

Using a similar thought process, one USGS source recommends using default values of alpha = 0.20 and a (smaller) beta = 0.10 (90% power) as defaults for power analysis to detect trends. The reasoning was that from the standpoint of conservation the consequences of sounding a false alarm (a Type I error) are small compared to the consequences of failing to detect a population crash (Type II error) (Eagle, P.C., J.P.Gibbs, S. Droege. 2000. Power Analysis of Wildlife Monitoring Programs: Exploring the trade-offs between survey design variables and sample size requirements. This document has been available online at: http://www.pwrc.usgs.gov/resshow/droege3rs/salpower.htm, but the most recent

**posting on that Website notes that the "MONITOR" calculator was temporarily removed for reconsideration and explained caution is advised due to "our growing awareness (and wariness) of the complexity of power analysis in general."**

**The point about the need for care and caution is well taken. There are many power and sample size and other statistical calculators of mostly unknown quality linked from websites such as:**

> **One attempt to compile links to various web pages that perform statistical calculations, including a very large number of this type, is found at www.statpages.net.**
>
> **There are diverse calculators on the UCLA website (http://www.stat.ucla.edu/calculators/powercalc). UCLA provides a large variety of sample size and power calculators for various one tail and two tail scenarios using normal, lognormal, binomial, exponential, and poisson distributions as well as correlation coefficients.**
>
> **The persistent who ask can still obtain the Patuxent Lab power analysis program "MONITOR" from Patuxent staff (see http://www.mp2-pwrc.usgs.gov/powcase/).**

**The default NPS suggestion is to limit beta (type II error) to 0.01 or 0.05 whenever possible. If beta is limited to 0.1 per the USGS suggestion above it insures that statistical power is at least 90%, and this may be acceptable in some situations, and will allow a smaller sample size than beta of 0.01. Limiting beta and not just alpha to low levels is one way to put more burden of proof on those trying to show "no difference," though often not as optimal as inequivalence testing.**

**Limiting false negatives (in addition to the more traditionally limited false positives) is one of the performance criteria (along with precision, systematic error (bias), sensitivity/detection limits, and specificity) needed to make sure performance based (PBMS) methods achieve data comparability (J. Diamond et al. 2001. Towards a definition of performance-based laboratory methods. A position paper of the National Water Quality Monitoring Council Methods and Comparability Board, Technical Report 01-02, Web:** http://water.usgs.gov/wicp/acwi/monitoring/nwqmc.

**When hypothesis testing is used, perhaps as one hint in one line of evidence, the hypothesis to be tested should be stated, and consideration should be given not only to whether or not there are statistical differences at defined levels of statistical power, but also whether or not the differences seen are of practical significance to the values to be protected or the desired future outcomes, or are otherwise large enough to influence potential management decisions (see discussion of decision rules in step III-B, and the discussion of Bio-inequivalence testing in step IV-C).**

**It is again stressed that some low level effects may be statistically significant but not biologically or ecologically significant. In other words they are de minimis (too small to be a problem) effects and might therefore be considered "bioequivalent." What effect sizes would be not only statistically significant but also biologically or ecologically significant? Once that is determined, what sample sizes**

**would be needed to see that effect size given sample variance? As suggested in a later step on monitoring design optimization, use variance estimates from past relatively high quality data for sample size calculation estimates.**

**Relatively simple descriptive statistics (bounded by uncertainty factors) along with a "multiple-lines-of evidence" approach can sometimes be used to estimate no or low effect levels. Describing the relationship between the distribution of exposure to contaminants and the distribution of effects is often appropriate for site assessments, whereas exposure-response modeling is most appropriate for the analysis of toxicity test data (Suter, G. W. II. 1996. Abuse of hypothesis testing statistics in ecological risk assessment. Human and Ecological Risk Assessment, Vol. 2 (2):331-347).**

**Again, there are many pitfalls related to using standard null hypothesis tests in field environmental monitoring or other observational studies. Such problems tend to be especially serious when one is looking at cause and effect inference and when beta is not limited (see Appendix IV.C. for more detail).**

## Consider Nonparametric Analyses

**As mentioned in section V-B, water quality and contaminants data is seldom normally distributed. This is why nonparametric methods are often helpful and have gained great favor in USGS water quality analyses. However, even nonparametric statistical methods are not "assumption-free" Nonparametric methods usually require random sampling and independent samples.**

**It should also be kept in mind that parametric statistics (based on squared deviations) are greatly affected by large values, and a single extreme value can largely control the estimate. On the other hand, nonparametric statistics (based on ranks) are insensitive to outliers. This can be an advantage when the outliers are mistakes in the data. There is a natural temptation to discard outliers, but they must be retained unless they can be shown to be mistakes. A potential disadvantage to using nonparametric methods in some situations (like in loading) is that nonparametric methods are not sensitive to a few large values. For example, high values of metals or high sediment concentrations in the water column may only be likely during infrequent high flow conditions. Although infrequent and appearing to be outliers, such high values may be the key to understanding load and transport dynamics. In other words, in some cases a few large values may legitimate and may constitute the most important data (Paul Geissler, USGS Patuxent Lab, Personal Communication 2002).**

**When asked what are the most common mistakes they see biologists and contaminants specialists making with statistics, some Ph.D. statisticians have answered that the most common mistake they see is that investigators do not attempt to determine what kind of underlying distribution the data comes from. Too often they don't even try to figure it out but simply blow and go with the wrong kind of statistics Often this takes the form of applying parametric statistics when the distribution was not normal, or transforming data and then creating data interpretation problems when back-transforming.**

One problem with the common sample size calculators on the net and in textbooks is that they tend to apply to parametric (normal distribution) applications only. Of the 550+ calculators listed as links on www.statpages.net in November 2003, none appeared to be specifically labeled a calculator for sample size for nonparametric applications. However, there were calculators for sample size or confidence intervals of a proportion (http://members.aol.com/johnp71/javastat.html#Power).

Although NP calculators of various types seem to be harder to find on the Internet, one can often multiply the sample size obtained from a parametric calculator by a constant to get an approximate sample size for NP work.

Using parametric power or sample size calculators on data that is not normally distributed, without multiplying the result times an efficiency constant, is wrong.

Relationships between parametric and NP test efficiencies are usually expressed as an asymptotic efficiency fractions. So if the fraction for a particular test was 0.9, it would mean that for large samples the parametric test requires only 90% as many cases to achieve the same power as its non-parametric counterpart. Conversely, the non-p test would require 1/0.9, or 1.11 times as many cases in this same case. The efficiency figures (at least the asymptotic efficiencies) have been worked out for a number of non-p test. Most of the famous non-p tests have pretty high efficiencies, so those needing to calculate Non Parametric Sample sizes often might want to do their homework and determine the conversion factors that would get them close in their particular scenarios. So, in summary I generally do power calculations for the parametric tests, and then just increase the sample sizes slightly if we're going to do a non-parametric test. (John C. Pezzullo, Biostatistics, Georgetown University Medical Center, Personal Communication, 2000).

For example, the nonparametric Mann-Whitney test has about 95% of the power-efficiency of its parametric analogue, the t-test (Siegel &Castellan, 1988, http://www.unep.ch/etu/archive/txt/mono5-3.txt). So if one used a parametric sample size/power to determine how many samples to take, one would have to multiply the result by 1/0.95 = 1.05 times to get the sample size for the specified amount of power using the Mann-Whitney test instead of the t test. Power-efficiency is discussed in Siegel's text (Siegel, Sidney. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Company. 1956. ISBN: 07-057348-4.).

**When the assumptions of the t test are badly violated, the Mann-Whitney test has more power than those tests that require assumptions of normality (J.H. Zar. 1984. Biostatistical analysis. Prentice Hall, N.J., page 141).**

**The Wilcoxen Sign Ranks test for two independent samples has about a 95% efficiency compared to a t test when the distribution is normal, (M.F. Triola and K. Karen Guardino. 1997. Elementary Statistics, 7 th ed. Addison-Wesley Press, ISBN #0201859203). So to get the sample size needed for this test, one could calculate the sample size needed for a T test while limiting alpha and beta both to stated levels and then multiply the result by 1.05.**

**For contrast, the nonparametric rank correlation test has about a 91% efficiency compared to parametric equivalents when the distribution is normal, (M.F. Triola and K. Karen Guardino. 1997. Elementary Statistics, 7 th ed. Addison-Wesley Press, ISBN #0201859203).**

**Sprinthall's textbook gives an alternative (rougher) rule-of-thumb guideline for Mann-Whitney sample size requirements, that for comparison of two samples, the Mann-Whitney test usually should only be used when each of the two samples is comprised of at least 9 values. If n (sample size) is less than 9 for each of the two samples, one has to find and use separate tables of critical U values (R.C. Sprinthall. 1987. Basic Statistical Analysis. Prentice Hall, NJ). However, it is probably better in most cases to calculate needed sample size by first calculating that which would be needed for the desired statistical power (beta) at given significance level (alpha) with a t test, then multiplying the result by 1.05 to get the sample size needed for the Mann-Whitney test, as described in the paragraph above.**

**Sprinthall also provides a similar (rough rule-of-thumb guideline) for sample size needed for the nonparametric (NP) Wilcoxen T test for paired (correlated) samples (a NP alternative to the parametric "paired T test." The Wilcoxen T test requires a sample size of at least 6-25 pairs of scores. When n is over 25, a separate equation must be used (R.C. Sprinthall. 1987. Basic Statistical Analysis. Prentice Hall, NJ).**

**Rule-of-thumb sample size guidelines have also been suggested for NP ANOVA procedures. The Kruskal Wallis H procedure for one way ANOVA should ordinarily have at least 3 groups and 6 subjects per sample. The Friedman ANOVA by ranks for three or more ordinal distributions with correlated selection ordinarily requires a minimum of "10 scores per column where there are 3 columns of ranked scores…With 4 columns of ranked scores, only 5 scores per column are necessary." Special tables must be consulted for smaller sample sizes (R.C. Sprinthall. 1987. Basic Statistical Analysis. Prentice Hall, NJ).**

**Many, if not most, experts would argue that power analyses should ideally be done before a study using the alpha value rather than retrospectively using the calculated p value. However, sometimes when one is looking back at someone else's data, the retrospective choice is the only one available.**

## Document Proposed Statistics

**The plan shall detail the statistics to be used in analyzing the data. The statistics required by regulators should be used when necessary to answer regulatory questions**

**The plan shall include an explanation of how the overall statistical sampling design will allow for inferences to be made to areas larger than those actually sampled. The plan shall identify specific populations to be monitored, and sampling units (http://science.nature.nps.gov/im/monitor/monplan.doc).**

## Data Quantity Objectives and Statistical Power

**The plan should include data QUANTITY objectives and requirements, i.e., what are the minimum sample sizes and sampling frequencies to ensure that sufficient data will be produced to be helpful in making previously identified decisions with adequate levels of statistical power, given what is known about parameter variance. Typically one uses variance from relatively high quality data historical data sets to estimate needed sample sizes. Sometimes however, pilot studies are needed to determine spatial and temporal variability characteristics.**

**Planners should keep in mind that long term monitoring often eventually produces large data sets but that useful data anticipated for any particular time period of concern seldom turns out to be 100% "complete" or acceptable from a quality standpoint after sampling is finished. Therefore, some "extra" samples typically need to be planned to make up for missing and/or low quality data. For more details on the data quality indicator "completeness", see step V-C of this document as well as the latest proposed EPA guidance in Section 4.2 of EPA 2001. Guidance on Data Quality Indicators (EPA QA/G-5i) at http://www.epa.gov/quality/ (Full text at http://on-linelearning.ca/idec4433/epaqaqc2000/g5i-prd.pdf).**

**The number of data points needed, and the frequency of sampling often depend on the requirements of State or federal regulatory authorities. Therefore, check with the applicable State or federal regulatory agency before proceeding with a sampling design and data quantify objectives .rather than relying on general national guidance such as the helpful but not always definitive (can be modified by individual states) guidance put out by EPA for TMDL programs, standards exceedances, and what is typically required to list waters as impaired (http://www.epa.gov/owow/tmdl/tmdl0103/text.html):**

> **"Larger data sets are particularly desirable when dealing with water quality criteria (WQC) with a fairly long duration factor (averaging period, like 30 days, 90 days, or a year). Hence, when making an assessment determination based on comparison of ambient data and other information to a numeric WQC expressed as an "average" concentration over a substantial period of time, a statement of a target number of samples may be appropriate. Still, the methodology should provide decision rules for concluding non-attainment even in cases where the target data quantity expectations are not met, but the available data and information indicate a reasonable likelihood**

> **of a WQC exceedance (e.g., available samples with major digressions from the criterion concentration, corroborating evidence from independent lines of evidence such as bio-surveys). However, small sample sets often provide sufficient information to support decisions to list waters because the frequency and/or magnitude of observed excursions and digressions are high enough to support a reliable impairment determination. Even a very small set of samples may be sufficient to indicate impairment, particularly when the duration/averaging periods of relevant WQC are quite short (an hour or less). When considering small numbers of samples, it is important to consider not only the absolute number of samples, but also the percentage of total samples, with concentrations higher than those specified in relevant WQC."**

**As mentioned in section II-C on questions to be answered, very specific questions driven by regulatory requirements will often drive how many samples are needed, frequency of sampling, location of sampling, etc. For example, consider the following question:**

> **Does the one-hour average concentration (based on a minimum of 5 samples per hour collected at least three times a month for one year, or other minimum sampling specified by the State) of copper from depth-composited water column samples at randomly chosen sites in a specified reach of river ever exceed the State water quality standard Criteria Maximum Concentration (CMC)?"**

## QC OBJECTIVES FOR DATA QUALITY INDICATORS:

### Difference between Data Quality Indicators (DQIs) and Measurement Quality Objectives (MQOs):

**In the past, some have listed DQIs to include precision, systematic error (bias), accuracy (which is systematic error (bias) uncertainty plus precision uncertainty), representativeness, comparability, completeness and sensitivity (PARCCS, see EPA guidance at http://www.epa.gov/quality/qs-docs/g5g-final.pdf).**

> **Note from Roy Irwin: One problem with using PARCCS in this manner is that if one properly understands "accuracy" to mean "uncertainty in accuracy" after factoring in both measurement precision systematic error (bias), then PARCCS leaves out an important QC concept, the control of systematic error (bias). If one means "uncertainty in accuracy" (after considering both precision and systematic error) when using the word accuracy, then one could use the acronym PARCCSS, with the last S standing for systematic error (bias). This is more in tune with NIST and ISO definitions (N. Taylor and C. E. Kuyatt. 1994. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results NIST Publication TN 1297 (http://physics.nist.gov/Document/tn1297.pdf). The other problem is that it is probably better to distinguish between data quality indicators**

**such as representativeness, comparability, completeness, and sensitivity versus more quantitative measurement quality objectives for precision, systematic error, and uncertainty in measurement accuracy. See sections discussing each of these concepts for more detail.**

**As of 2001, federal agencies such as EPA, USGS, NOAA, and methods standardization groups such as the Methods and Data Comparability Board (MDCB) Accreditation Workgroup of the federal (interagency) National Water Quality Monitoring Council have all endorsed the need for validated methods and concise, achievable performance criteria for data quality objectives and data quality indicators (J. Diamond et al. 2001. Towards a definition of performance-based laboratory methods. A position paper of the National Water Quality Monitoring Council Methods and Comparability Board, Technical Report 01-02, Web:** http://water.usgs.gov/wicp/acwi/monitoring/nwqmc. **The NPS endorses these same concepts for either qualitative or semi-quantitative data quality indicators, including representativeness, comparability, sensitivity, and completeness. These are discussed in turn:**

## IV-D. Data Representativeness, A QC Data Quality Indicator

**A key part of quality control is to make sure that sample observations are really "representative" of the target population being sampled and/or the actual condition being measured or estimated. Decisions related to randomization and other study design details are often important in insuring representativeness.**

**Representativeness is a data quality indicator that historically has been controlled qualitatively. For a more detailed explanation of how to assess representativeness, see section 4.1 of latest proposed EPA guidance [EPA 2001. Guidance on Data Quality Indicators (EPA QA/G-5i) at http://www.epa.gov/quality). For more detailed discussions of related variability issues, see EPA's summaries at http://www.epa.gov/emap/html/pubs/docs/resdocs/ecol_ind.pdf and NPS I&M Guidance at http://science.nature.nps.gov/im/monitor/epacrit.htm#phase3.**

**A useful document on the internet that gives guidance on survey design vs. representativeness and sample units is EPA. 2002. Guidance on Choosing a Sampling Design for Environmental Data Collection (QA/G-5S). EPA/240/R-02/005 (http://www.epa.gov/quality/qs-docs/g5s-final.pdf).**

## IV-D.1 The Need to Understand Patterns of Variability

**A study of flawed/failed monitoring projects revealed that many problems could have been avoided if the project designers had had a clear understanding of the temporal and spatial variability and "magnitudes of response" of parameters measured (L.M. Reid. 2001, The epidemiology of monitoring. Jour. Amer. Water Resources Assn. 37(4): 815-819).**

**Unless one understands the temporal and spatial variability in various potential strata, it is difficult to designate defensible strata in stratified random**

**sampling (see section VI-D.1 below). Strata should have relatively homogenous variability characteristics.**

**It is accordingly important to avoid oversimplified or misleading assumptions related to variability. Incorrect or only partly correct assumptions related to variability in time and space frequently hinder getting representative samples.**

**For example, although many water quality specialists seem to understand parameters like oxygen and pH have strong daily and seasonal variability, it is probably less well known that concentrations of metals in the water column and sediments of various streams can also vary seasonally and can have very different variability characteristics in different types of microhabitats (study design strata).**

**Likewise, it is commonly known that water column values for conductivity, turbidity, suspended solids, and metals often vary with flow volume, but less commonly recognized that water column concentrations of, manganese, arsenic, cadmium, copper, and zinc, can vary up to 500% over a 24-hour period IRRESPECTIVE OF CHANGES IN STREAM FLOW (David Nimick and Aida Farag, USGS, Personal Communication, 2001).**

**Fluctuations in the concentrations of metals in solution are related to pH because pH regulates dissolution, precipitation, sorption and desorption reactions, Most metals will be more water soluble and therefore have a higher concentration in the water column as pH goes down (becomes more acidic). However, arsenic, methyl mercury, chromium (chromate anion), phosphorus (phosphate anion), selenium (selenate anion), and manganese (manganate anion) tend to behave differently (see appendix IV-D.1 for more detail).**

**In many water bodies, swings of pH are indirectly caused by photosynthesis during the day and plant respiration at night. Aquatic plants (including but not limited to phytoplankton, periphyton, and higher plants) tend remove carbon dioxide from the water column during the day as they photosynthesize, resulting in lower concentrations of carbonic acid and therefore higher pHs. At night, plants respire and move carbon dioxide into the water column, resulting in more carbonic acid in the adjacent water column and therefore indirectly lowering pH. These typical diurnal fluctuations in pH play a role in metals concentrations. As pH goes up, water column concentrations of most metals (other than arsenic and the other compounds listed above, see appendix IV-D.1) go down. The reverse is true when pH goes down. In other words, water column concentrations of most positively charged (cat ion) metals vary strongly and inversely with pH.**

**In a study of a freshwater stream in the west, metals appeared to be follow the generally recognized pattern, being mobilized from a scrape-able >heterogeneous surface layer (a layer of abiotic substances as well as >periphyton and non-algal biofilms) into the dissolved phase as pH goes >down, and then move the opposite way, out of the water column as pH goes >up. This tends to cause water column metal concentrations to be lower during daylight hours and higher at night (J. Morris et al. 2003, Profiling Synthetic Biofilm using Miniaturized Ion-selective electrodes. Platform Presentation at the Rocky Mountain SETAC meeting, April 16, 2003, Denver, Colorado).**

**More recent analyses by Jeff Morris show that zinc is removed from the water column during the daylight hours (due to the light-driven processes), but it is not released back into the water column at night. Thus, the biofilm (or more likely, the Mn-oxide crust under the biofilm) appears to act as a "permanent" sink for the zinc (Joe Meyer, University of Wyoming, Personal Communication, 2003).**

**Sulfate reducing bacteria, which can help metals change form (and end up immobilized in sediments as metal sulfides in some cases or mobilized in the case of the formation of bio-accumulating methyl mercury) can also play a role in changing metals concentrations in water columns. We usually think of most sulfate reduction of metals to sulfide forms occurring in anaerobic sediments.**

**In free flowing rivers with an abundance of oxygen and relatively little sulfate, the water column is often aerobic but anaerobic or sulfate reducing conditions can nevertheless form just inside of water column/sediment "edges." This would include edges between the water column and the insides of biofilms, Cyanobacteria mats, or anaerobic sediments.**

> **For a discussion of Cyanobacteria mats see A. Teske,N. B. Ramsing, K. Habicht,M. Fukui, J. Ver, B. Jørgensen, and Y. Cohen. 1998. Sulfate-Reducing Bacteria and Their Activities in Cyanobacterial Mats of Solar Lake (Sinai, Egypt), Applied & Environmental Microbiology Vol. 64 (8) 2943–2951, American Society for Microbiology.**

**Anaerobic (sufate reducing) conditions can even form "in "an envelope of water surrounding a particle of organic matter that changes from an aerobic environment to an anaerobic environment over time (Nancy Simon, USGS, Personal Communication, 2003).**

**The main role biofilm could play in this situation is to form a sharp microhabitat edge by creating a boundary between the zones of anaerobic and aerobic respiration due to the slow diffusion of oxygen through relatively thick biofilm mats as well as high levels of aerobic respiration by algae, microbes and other organisms in the mat. Thick biofilm mats are usually formed in areas with high nutrient inputs and/or warmer temperatures, but can form even in Montana, a relatively cold climate. In a productive system, there would be plenty of organic matter for anaerobic microbes to chew on and once they depleted the nitrate, manganese and iron, they would begin using sulfate as a terminal electron acceptor (different species of microbes for each acceptor, of course) and start producing sulfide that could bind up metals such as zinc (Jeff Morris, University of Wyoming, Personal Communication, 2003).**

**These details are relevant to environmental monitoring as they help explain why water column concentrations of metals can vary up and down diurnally or seasonally.**

**In most streams and many other water bodies (perhaps especially in shallow streams with little cover, strong photosynthetic processes, and weak buffering), one might have a difficult time tracking long term trends of water column concentrations of metals unless the data was normalized to pH or at least sampling was standardized related to a constant amount of time before or after sunrise. Since pH can also vary seasonally,  and since temperature can also play a role in water solubility of metals and other substances, testing for trends only in seasonally adjusted manner (for example, using the seasonal Kendall test) may also be appropriate.**

**Until standard methods are developed to normalize water column concentrations of metals to pH, those planning monitoring should at least be aware of the issues and should also consider trying to standardize the time of collections compared to hours from sunrise and/or season.**

**Sediment concentrations of metals in rivers have been know to vary according to season (Nimmo, D.R., Willox, M.J., Lafrancois, T.D., Chapman, P.L., Brinkman, S.F., and Greene, J.C., 1998, Effects of metal mining and milling on boundary waters of Yellowstone National Park: Environmental Management, v. 22, p. 913-926) and to be controlled partly by microhabitat type (riffles, runs, backwater pools, etc.).**

**In Soda Butte Creek near Yellowstone NP, variability in metals also varied by microhabitat. At any given time, variability tended to be less in low gradient riffles than in back water pools and was highest in seasonally submerged attached bars (Ladd, Scott, Marcus, W. Andrew, and Cherry, Steven, 1998, Trace metal segrega-tion within morphologic units: *Environmental Geology and Water Sciences,* 36(3/4):195-206).**

**So in this case if one was looking for long term trends in metals concentrations, one might choose to look in low gradient riffles rather than in back water pools, to minimize variability and make it easier to pick out trends with a reasonable number of samples. Standardizing sampling periods can also be helpful.**

**For example, metals concentrations in sediments measured each August over six years along the length of Soda Butte Creek remained approximately the same, despite the occurrence of 15, 50 and 100-year flood between sampling periods( Marcus, W. Andrew, Meyer, Grant A., Nimmo, D.R, 2001, Geomorphic control on long-term persistence of mining impacts, Soda Butte Creek, Yellowstone National Park. Geology 29(4):355-358, http://geography.uoregon.edu/amarcus/Publications/Marcus_et_al_2001_Geology.pdf).  However, metal concentrations at the same sampling sites varied considerably from month to month over one year, although concentrations returned to approximately the same levels at at given time of year (e.g., August concentrations were approximately the same in each year sampled) (Marcus, W. Andrew, Ladd, Scott, Crotteau, Michael, 1996, Channel morphology and copper concentrations in stream bed sediments: in Nelson, John D. et al. (eds.), *Tailings and Mine Waste '96*, Balkema Press, Rotterdam, p. 421-430).**

**Given the characteristics of metals variability documented by Marcus (above) and the fact that metals move from sediments to the water column and back with changes in temperature and pH, it is not surprising that concentrations and variability in metals concentrations could change with season. Again, those designing monitoring for long term trends of metals in sediment might wish to consider monitoring in habitats (like low gradient riffles) with low baseline variability and restrict monitoring only to certain seasons.**

### IV-D.2: Include Some Aspect of Randomization

**Usually, some aspect of randomization (stratified random sampling, spatially balanced random sampling, simple random sampling, etc.) needs to be incorporated in the survey design to insure equal probability of sample selection and therefore help insure that the samples selected are representative of the population being sampled and that statistical inferences from the sample to the larger population can be made.**

**Social scientists learned 50 years ago that judgment sampling was inaccurate. The great expense of monitoring natural resources, combined with the stakes of coming up with the wrong answer, suggests we cannot waste another 20 years and large amounts of money waiting for those monitoring the environment to relearn this lesson (D. Edwards. 1998. Issues and themes for natural resource trend and change detection. Ecological Applications 8(2):323-325, http://www.esa.org/sbi/sbipubs.htm).**

**Addressing this issue, I&M guidance states: "Some sort of probability sample should always be taken to avoid systematic error (bias) in sampling locations. Conceptually, the target population (usually the entire park) is divided into sampling units such that every point in the park is included in a sampling unit, but not in more than one. The sampling design is used to select a probabilistic sample of the sampling units. As a result, statistical estimates of population attributes can be produced with an estimate of their reliability. Probability samples occur when each unit in the target population has a known, non-zero probability of being included in the sample, and always include a random component (such as a systematic sample with a random start). The credibility of data that are not collected using these principles is easily undermined" (http://science.nature.nps.gov/im/monitor/#Design).**

**One bonus reason for incorporating random sampling in the study design is that it might facilitate future investigations or statistical tests of subsets of data to provide insight concerning future (but currently unknown) questions.**

**Although some type of randomization is typically necessary to insure representativeness, it is wrong to assume that pure, non-stratified random sampling is always the only way to go. Other aspects that typically need to be considered include the following:**

1) **If the issue is what is the concentration of water column chemicals coming down a river, and one desires to obtain data comparability with USGS data, one must typically use USGS methods of compositing samples**

through vertical and horizontal profiles according to USGS guidance (http://water.usgs.gov/owq/FieldManual).

2) Often budget and other practical considerations result in the need to change questions so that one can utilize stratified random sampling rather than pure random sampling.
3) Typically networks should consider not only simple randomization, but also spatially balanced randomized sampling. EPA has suggested a unified strategy for selecting spatially-balanced probability samples of natural resources, noting that "The spatial distribution of a natural resource is an important consideration in designing an efficient survey or monitoring program for the resource. Generally, sample sites that are spatially-balanced, that is, more or less evenly dispersed over the extent of the resource, will be more efficient than simple random sampling" (D. Stevens and A. Olsen, Spatially-Balanced Sampling of Natural Resources in the Presence of Frame Imperfections, http://www.orst.edu/dept/statistics/epa_program/docs/spatial_balance_imperfect_frame.pdf). For additional detail on spatially balanced designs, see appendix IV-D.2.

Alaska NPS staff have developed a "NPS Sampling toolkit extension (SamplePAK)." This includes tools for creating, manipulating and analyzing regular (rectangular) sampling grids. It can be useful in developing sampling strategies, analyzing sampling densities, and along with the AlaskaPak toolkit's random selection tools, can be used to stratify and randomize sampling approaches (http://www.nps.gov/akso/gis/av31/sampak.htm).

Typical Park Service qualitative data quality objectives:

The plan shall document how the data being collected is "representative" of the target population being studied. Again, the larger population is the larger population about which statistical inferences are to be made, after considering study boundaries, population (true) temporal variability, and population (true) spatial variability.

Probability samples that in some way make use of randomization should be used when the goal is to make probability statements about the quality of estimates or hypothesis tests that are derived from the resultant data (EPA. 2000. Guidance for Data Quality Assessment Practical Methods for Data Analysis, EPA QA/G-9, EPA No. EPA/600/R-96/084, http://www.epa.gov/quality/qs-docs/g9-final.pdf).

Rationales for choosing judgmental versus probabilistic designs, and for choosing probabilistic options such as simple random sampling versus more advanced sampling designs (such as ranked set sampling, adaptive cluster sampling, see EPA. 2002. Guidance on Choosing a

**Sampling Design for Environmental Data Collection (QA/G-5S). EPA/240/R-02/005 (http://www.epa.gov/quality/qs-docs/g5s-final.pdf) should be provided in the plan.**

**If no randomization is done, one typically needs to document the rationale as to why samples are nevertheless "representative" and appropriate in the plan.**

**As suggested in I&M guidance, the plan should "describe the approach used to determine where sampling will occur for each vital sign, including justification for collocating or not collocating sampling for various vital signs. Provide justification for the attributes used to stratify the park, attributes such as cost of access, terrain features such as elevation and slope, and a soils or vegetation map" (http://science.nature.nps.gov/im/monitor/monplan.doc).**

**The plan should identify practical survey design constraints on collecting data that would limit the ability to representatively sample potential target populations.**

**Decisions about sampling only certain sample unit categories (strata in stratified-sampling schemes) should not be done in a vacuum. Unless the differences in variability and typical quantities in various spatial and temporal classes (the potential strata) are well understood and can be documented, or unless the project planners are willing to restrict all inferences to these groupings, these classes should not be selected as strata.**

**If rare species and other values to be protected are only found in certain types of rare microhabitats, then the sampling strata should be designed to include representative sampling in those rare microhabitats. In other words, transects and small plots or other sampling schemes likely to miss the relatively rare strata or microhabitats of concern should not be used.**

**Destructive sampling is discouraged unless unavoidable. If it is needed, special considerations and precautions to be taken should be documented in the plan.**

**For variables known to respond strongly to normalizing variables, the normalizing values should be taken into account in study design and should be measured at the same time as the variable being observed, and reported as metadata (see appendix IV-D for details).**

## IV-F. Final Monitoring Design Optimization Steps (Optional In Monitoring Plans but a Highly Desirable for Aquatic Projects).

**Before the plan is finalized and final peer review is obtained, the key monitoring plan authors and technical experts need to document final decisions on monitoring plan optimization.**

**Due to limitations of funding, logistics, and feasibility, typically monitoring networks will not be able to afford to sample all parameters at all sites as frequently as would be optimal. Other general considerations include:**

1. **Lack of ability to detect changes in face of high natural variability at pristine sites,**
2. **Difficulty in monitoring effectively (or find trends or answer questions) due to measurement uncertainty or model uncertainty being too high,**
3. **Duplication of Efforts: Other groups may already be monitoring similar parameters at the same or nearby sites, so there is less urgency for the NPS to do so.**

**Although generic VS guidance (Outline for Vital Signs Monitoring Plans, 2003, http://science.nature.nps.gov/im/monitor/docs/monplan.doc) does not require it, for aquatic projects, WRD suggests it is highly desirable to document the final monitoring plan optimization steps considered in fine tuning and optimizing the monitoring plan. Typically optimization steps to be considered and documented should include:**

1. **General Monitoring Design Optimization Steps (Discussed in Section IV-F.1, below).**
2. **Summary of Steps Taken (if any) To Minimize Measurement Uncertainty, as discussed briefly in Section IV-F.2.**
3. **Summary of Steps Taken (if any) To Minimize Model and Study Design Uncertainty Discussed briefly in Section IV-F.3,**
4. **Brief Description Of Plan To Implement Pilot Scale Monitoring (If Applicable) in Section IV-F.4.**
5. **Description Of Who Will Revise The Plan Following Pilot Scale Monitoring And When Long-Term Monitoring Will Begin (If Applicable) in Section IV-F.5, below,**

### IV-F.1 General Monitoring Design Optimization Steps

**Once the team has discussed all the previous planning steps experience has shown that changes are almost always made. What changes in the plan are necessary to optimize the monitoring design?**

**Typical NPS study design optimization objectives:**

### IV-F.1.1 Review and Reconsider Sampling Sites, Sample Sizes and Sampling Frequencies

**At this point, the team should reconsider some issues first considered in a less detailed way in step V (study design, statistics, power). What data is needed to answer questions given parameter variability? What is known about variability structure and variance estimates for each proposed sampling parameter in each proposed strata should be reviewed. Power analyses should then be done to determine sample sizes needed to answer each proposed question given the decision rules and the effect sizes that need to be detected. This needs to be repeated for each question in relationship to proposed variables to be monitored in each proposed sample unit or strata.**

> **Note: the goal here is to optimize sampling intervals in time and space and to provide information needed in the next steps (determining which questions cannot be answered within budget limitations and making cost/benefit study design decisions). Is variability between years greater than seasonal variability? Is diurnal variability and/or variability at pristine sites so high that a prohibitive number of samples would have to be taken to determine if decision rule thresholds have been surpassed? Is the variability of contaminant concentrations between individual fish of the same size greater than seasonal variability? How do variance and mean, median, or confidence interval values differ in various subdivisions of time and space? What are the differences in these estimates between pristine sites and impacted sites? The answers to these types of questions can greatly impact study design, and lack of knowledge about these issues often drives one to do pilot studies and/or extensive analyses of other regional datasets.**
>
> **Practical approaches are suggested:**
>
>> **For example: Monitoring groups often do not have enough funding to answer all questions, sample all sites, and sample as often as one would like to. A practical way to use sample size calculators is to use variance estimates from relatively recent data and high quality data sets (rather than all data) in various categories, for example high flow vs. low flow or summer vs. winter. One might decide that recent USGS data is among the better (based on QA/QC and other considerations) data available. One might then decide to use only these data sets to get variance estimates for various categories of data. If the variability of a certain parameter was very high even at pristine sites, and sample size calculators suggested a very large (too costly) number of samples were needed to detect a decision rule change (say a 20% effect size for example), it might be justification for deciding to study other variables and/or other location strata or simply to address other questions. If no data from the exact stream in question exist, one**

**typically needs to determine if reasonable variance estimates can be obtained from similar sized streams in the same general area (same general region, altitude, stream order, climate, geology, etc.)? If not, one might have to get pilot scale data to obtain reliable variance estimates needed for required sample size and required statistical power calculations (J. Loftis, 2001. Case study on the design and implementation of a water quality network: statistical considerations, Invited Lecture Presented to the NPS Water Quality Vital Signs Workshop, November 29, 2001, Water Resources Division, NPS, Fort Collins, CO).**

**Where does one get recent high quality data to estimate temporal and spatial variability? For freshwater, the answer is often USGS. For estuarine waters, one source is NOAA's NERR program (http://inlet.geol.sc.edu/cdmoweb/overview.html).**

**Among the basic issues to be revisited at this stage include data quality objectives, methods to be used, and the following sampling design basics (http://www.epa.gov/quality/qs-docs/r5-final.pdf):**

**The types and numbers of samples required**

**The design of the sampling network,**

**The sampling locations and frequencies,**

**Sample matrices,**

**Measurement parameters of interest, and**

**The rationale for the design.**

**Section V of the detailed monitoring plan should changed according to any changes decided at this stage, and Section IV-F.1 of the detailed study plan should include a summary of any study-optimization phase final changes in the overall statistical design, who will do the analyses, and how often.**

### IV-F.1.2 Reconsider Screening Vs. Definitive Methods

**At this stage, planners often determine that funds are too short to do everything they wanted in the most rigorous way. Given project needs, funding limitations, and information content of data are less expensive screening methods sufficient? This is a good time to reconsider whether or not the SOP and QA/QC methods chosen produce the most bang for the buck given funding limitations and data quality needs. Definitive methods are not always more optimal in producing maximum information-content when compared to screening methods. Rather than giving up**

**trying to answer a question because of cost, one should often re-examine whether or not there are screening methods that would answer and still provide adequate information quality. In biological sampling of invertebrates, is it really necessary to identify every specimen to species level, or would family level identifications give "taxonomic sufficiency" for the task at hand? The State of NJ and TVA both decided that using the family level for taxonomic sufficiency provided more information per time and funding expended.**

**In biomonitoring, sometimes using a relatively simple state multi-metric method provides good information content and data comparability compared to using more expensive methods. In chemical analyses, is the most definitive and expensive method really needed, or would less expensive methods give adequate MQO performance to meet QA/QC objectives? If the concentrations of the analyte are high in all samples, do you really need that more expensive "low detection limit" method? In the trade off between using screening methods to cover more time and space vs. doing a more definitive method at fewer places and fewer times, which choice would provide the most useful information and the most information content relevant to the question? See the definition of useful data and consider the following thoughts from an EPA guidance document:**

> **If the goal is obtain a defensible site assessment that reflects the true site condition, there are scenarios where fewer higher quality data points leads to lower information value compared to many (lower quality) data points. Analyzing samples using a highly accurate method can be very expensive, so often the hard truth is that budget constraints frequently limit the number of samples used to determine the presence and degree of contamination. It should also be kept in mind that environmental decisions can be especially susceptible to error in site cleanup situations because the major source of decision uncertainty (as much as 90% or more by some estimates) is due to sampling variability as a direct consequence of the true heterogeneity of environmental matrices. These situations drive home the concept that sometimes highly accurate and QA-documented (i.e., high quality) data points may actually form a poor quality data set that produces misleading conclusions and erroneous project decisions. Part of the DQO process is to ensure the data are "effective" for the purpose at hand, not necessarily "perfect." For more information, see definition of useful data and source document for most of the thoughts in this paragraph (D. M. Crumbling. 2001. Applying the concept of effective data to Environmental analyses for contaminated sites, Current Perspectives in Site Remediation and Monitoring. EPA 542-R-01-013, www.clu-in.org, choose publications and studio, then Characterization and Monitoring).**

### IV-F.1.3 Drop Monitoring Beyond Budget or Already Being Done By Other Groups

**Is another group already doing the monitoring. Is a neighbor group already monitoring streams impacted by air pollution? Who can we partner with to reduce or eliminate costs?**

**Following team consideration of the issues listed above, the plan should detail the questions, parameters, strata, and sample units that could not be addressed or monitored due to funding limitations. Variance characteristics and sample sizes needed to detect decision rule thresholds may simply make it impractical to measure certain proposed vital signs in the boundaries of time and space originally envisioned.**

**The cost/benefit or cost/effectiveness of the overall study design choices, including sampling and analysis options, should be discussed by the team and then finalized in the plan, to reflect the goal of optimizing the study to best answer with questions within funding limitations. Any other "basic" monitoring or sampling design changes deemed appropriate and desirable should also be discussed and summarized.**

**A practical approach is recommended:**

> **A first step of the design process is to quantify, as far as is possible, the information required by Park management, regulatory agencies, and the public. A second step is to quantify the information the monitoring system is capable of producing given funding levels. If after completion of the first two steps, there are differences in the information expectations and the ability of monitoring to produce information, then the management strategy, the monitoring budget, the monitoring system design, and/or the law itself may have to be examined and reformulated. This may seem to be over stated; however, it is simply a waste of money to monitor without a clear relationship between the information to be produced and its use within the management agency's (and or regulatory agency) decision-making process! It is the purpose of water quality monitoring design to ensure that monitoring is well integrated into the information needs of water quality management (R. Ward, 2001. A systems perspective and design framework for water quality monitoring, Invited Lecture Presented to the NPS Water Quality Vital Signs Workshop, November 29, 2001, Water Resources Division, NPS, Fort Collins).**

### IV-F.2 Summary of Steps Taken (if any) To Minimize Measurement Uncertainty

**Uncertainty about a single data point reflects an estimate of our lack of confidence in the overall accuracy of the value obtained for that one single data point after factoring in QC results for precision (specified as either reproducibility or repeatability), QC results for systematic error (bias), and any other obvious contributors to uncertainty (such as uncertainty in the certified reference material**

**used to estimate systematic error). Measurement systematic error (bias) is a systematic (usually consistently low or high) error.**

**Uncertainty about a summary statistic such as a mean is a different concept, one that includes uncertainty in the true variability (heterogeneity) in the samples, not just uncertainty about a single data point. Uncertainty about a summary statistic such as a mean is a concept addressed in more detail in subsequent discussions and a concept usually best addressed by using a measurement uncertainty-corrected confidence interval that addresses both measurement uncertainty (the halo of uncertainty about each data point) as well as uncertainty related to true variability (sample heterogeneity) of the values being measured (the halo of uncertainty about the mean of raw values uncorrected for measurement uncertainty). Uncertainty about a mean is usually expressed as a standard statistical confidence interval, such as a  t distribution confidence interval or a nonparametric confidence interval.**

**Returning to the subject of uncertainty about an individual data point, the scientific community and engineering community have reached some global consensus points on certain issues. For example, the U.S. (NIST) and international (ISO) agencies charged with standardizing measurement methods and terminology used in science and engineering have reached consensus on measurement basics. These include:**

1. **No measurement is perfect. Each is an approximation, and**
2. **Individual measurement data points are not complete unless accompanied by a statement about the uncertainty of that approximation.**

**These are hardly new concepts. Since's 1963 paper (C. Eisenhart. 1963. Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems, *J. Res. Natl. Bur. Stand.* 67C, 161-187), it has been broadly accepted that measurement uncertainty is a broader term than just variability, and that the concept should include: (1) a measurement process requires statistical control; (2) statistical control implies control of both reproducibility and repeatability; and (3) a measurement result requires an associated statement of uncertainty that includes any possible source of systematic error (bias) (http://nvl.nist.gov/pub/nistpubs/sp958-lide/html/129-131.html). One of Eisenhart's often quoted statements was that "Until a measurement system is in a state of statistical control, it cannot be believed in any logical sense that it is measuring anything at all."**

**Furthermore, the scientific community has known since the 1930's that measurement processes need to be controlled for both precision and systematic error (bias) (Newman, M.C. 1995. Quantitative Methods in Aquatic Ecotoxicology, Lewis Publishers, Boca Raton, FL., p. 282).**

**However, some of this "news" from the 1930's seems to have not fully filtered down to many water quality, contaminants, and biological monitoring specialists in federal and state agencies and academia.**

**NIST has standardized with ISO simple algebra equations for estimating measurement uncertainty (N. Taylor and C. E. Kuyatt. 1994. Guidelines for**

**Evaluating and Expressing the Uncertainty of NIST Measurement Results NIST Publication TN 1297 (http://physics.nist.gov/Document/tn1297.pdf). The equations include sum of squares, with the squares being squares of standard deviations of precision, systematic error, imperfections in certified reference materials, and any other major contributors to uncertainty. One then takes the square root of the this sum of squares and multiplies it by the middle t value if sample size is less than 30 to come up with a kind of pooled square root to express measurement uncertainty**

**For the situation regrettably common to water quality, biological, and contaminants monitoring, where historically data has not been adjusted for best estimates of average systematic error (bias), NIST has also provided guidance on how to calculate asymmetrical "level of confidence" intervals about single data points. For values not adjusted for systematic error this typically results in "SUM*U* level of confidence" intervals (not confidence intervals on a mean) applicable to each individual measured value (S.D. Phillips, K.R. Eberhardt, and B. Parry. 1997. Guidelines for Expressing the Uncertainty of Measurement Results Containing Uncorrected Bias. Journal of Research of the National Institute of Standards and Technology, Volume 102, Number 5, September–October 1997, http://nvl.nist.gov/pub/nistpubs/jres/102/5/j25phi.pdf).**

**Whether symmetrical or not, measurement uncertainty intervals based on these NIST/ISO standardized equations are easy to calculate. In choosing which vital signs to monitor, and in rethinking the overall design during monitoring plan optimization steps, planners should take another look at the measurement uncertainty of various vital sign/measures and consider throwing out those measures whose measurement uncertainty is so high that it greatly compromises the possibility of using those vital signs to answer questions or determine if trends are occurring.**

**Step-by-steps for calculating measurement uncertainty and other details related to measurement uncertainty are provided in Appendix IV-F.2.**

### IV-F.3 Summary of Steps Taken (if any) To Minimize Model and Study Design Uncertainty

### Model Uncertainty

**Planning networks first looked at uncertainty in model components (particularly resources to be protected vs. potential measures/vital signs) in section II-C, above. Qualitative (Type B, expert judgement) uncertainty was estimated as either high, medium, or low at that stage. Now that networks have progressed to this later monitoring plan optimization stage, they are often trying to eliminate vital signs and and/or sites due to lack of funding. At this stage, it is suggested that networks reconsider dropping measures that do not have a strong relationship to values to be protected in terms of desired future conditions (DFCs).**

**Some examples of connections that might be shown to have a low degree of uncertainty might include:**

**Proposed Vital Sign: pH reading in the water.**

> > **Desired Future Condition. Relief from acid mine drainage effects as shown by pH changing to vary within acceptable natural ranges, and as confirmed by another proposed vital sign, biodiversity of aquatic macro-invertebrates.**

> **Proposed Vital Sign: Total Nitrate-Nitrogen Concentration in the water column.**

> > **Desired Future Condition: Relief from unnatural eutrophication conditions, shown to be nitrate limited in the area.**

**An example of a model connection that might be shown to have an even higher degree of conceptual model (connection) uncertainty might be the following:**

> **Proposed Vital Sign: Different observer's best estimates of "percent embededness in cobbles in stream bottom sediments" based on qualitative habitat observations.**

> > **Desired Future Condition: A restored, healthy population of topminnows of the genus <u>Fundulus</u>. The model uncertainty of this connection might be shown to be high or unknown since it is known that topminnows mostly hang out at the surface of the water and eat surface insects rather than interacting with cobbles in bottom sediments. Furthermore, in this hypothetical example, let's assume that no one in region has documented any kind of relationship between percent embededness of cobbles and the well being of topminnow populations. Lastly, for percent embededness, another type of uncertainty, measurement uncertainty, has been shown to be very high due to high reproducibility imprecision between observers. While not model uncertainty, this represents another disadvantage of picking this variable as a vital sign, another disadvantage related to one's ability to show deleterious trends with reasonable sample sizes, and therefore with reasonable cost.**

**There are statistical ways to estimate model uncertainty, but these tend to be complex, so they are discussed in appendix IV-F.3.**

**Study Design Uncertainty**

**At this later monitoring plan optimization step, it is suggested that monitoring networks also take a final look at study design uncertainty and consider dropping vital signs and stations where the uncertainty is unacceptably high.**

**For example, sampling uncertainty related to lack of representativeness of the sample, will often have an even larger effect on the results and conclusions than measurement uncertainty. This is why one must take pains to make sure that the**

**sample being taken is randomly selected from the population of interest. For more information on survey design, see http://science.nature.nps.gov/im/monitor/#Design.**

**So at the monitoring plan optimization stage, consider throwing out variables whose representativeness (see section IV-D) to the identified population being sampled is shaky (has not or cannot be controlled in optimal ways). Also consider throwing out vital signs and sites where patterns of variability (in time and space, see sections IV-D.1 and appendix IV-F.3) are not understood well enough to be able to defend designations of strata or other sampling design basics.**

**If one has conscientiously followed the outline herein and gone through the various QA/QC steps, many sources of uncertainty should have been controlled or at least minimized. However, by this stage in the planning process, planners should also have a feel for which vital sign/measures and sampling components have greater degrees of uncertainty than other. At this stage, it is suggested that planners consider dropping those components whose uncertainty is either unacceptable or those components whose uncertainty is simply comparatively high.**

### IV-F.4. Brief Description of Plan to Implement Pilot Scale Monitoring (If Applicable)

**For those vital signs/measures whose variability in time and space in pristine environments is not well understood, one has two options:**

1. **Throw them out and monitor something else better understood which does not require pilot scale monitoring.**
2. **For those parameters considered too important to throw out, initiate pilot scale monitoring of at least one or two years to better refine estimates of spatial and temporal variability, study design feasibility, required samples sizes, and data completeness.**

**Whether officially considered pilot stage or not, at least some of the initial data collected should be analyzed very soon after information is first collected, to allow for adjustments to be made if problems arise. A study of flawed/failed monitoring projects revealed that many of the problems could have been avoided if this had been done (L.M. Reid. 2001, The epidemiology of monitoring. Jour. Amer. Water Resources Assn. 37(4): 815-819).**

> **Note: In a related thought, a NAWQA veteran concluded that “what the monitoring program in Cycle I (1992-2001) perhaps could have done better was ensure sufficient people were in place sooner to analyze some of the initial QC data obtained at the very beginning of each new three-year sampling period so that if data-quality problems appeared, they could be identified and corrected earlier rather later. Also the periodic national aggregation and analysis of QC data which occurred at the end of every three year sampling period could be made a higher priority to ensure that results would be available**

> **earlier to those analyzing the water-quality data obtained for that period" (Michael T. Koterba, USGS, Personal Communication 2002).**

### IV-F.5. Description Of Who Will Revise The Plan Following Pilot Scale Monitoring And When Long-Term Monitoring Will Begin (If Applicable).

**The study plan should be revised based on any lessons learned during the pilot scale phase before long-term monitoring is implemented. Who will do these revisions, and if they are planned, when will more definitive long term monitoring begin?**

# V. SAMPLING PROTOCOLS

**As suggested in generic VS guidance (K.L. Oakley, L.P. Thomas, and S.G. Fancy, 2003. Wildlife Society Bullet 31(4), reprint at http://science.nature.nps.gov/im/monitor/protocols/ProtocolGuidelines.doc, all sampling protocols will include three basic sections:**

**A. Protocol Narrative**
**B. Protocol Standard Operating Procedures (SOPs), and**
**C. Protocol Supplementary Materials.**

**For aquatic projects (and any other projects where quality control is important), Quality Control details will ordinarily be included in a QC SOP as part of each protocol. Each QC SOP should include and details of data representativeness not adequately covered in the monitoring plan plus how quality control will be accomplished for each of the following QC basics (summarized in sections V-A to V-I):**

## V-A. Data Comparability, A QC Data Quality Indicator

### V-A.1 Standard Operating Procedures (SOPs) and other Standard Methods

**Listing standard operating procedures (SOPs) and other method details (associated with "protocols" in VS guidance) is a Quality Assurance (QA) basic. It is also necessary before one can determine whether or not data comparability, a quality control (QC) data quality indicator basic, has been assured.**

**Data comparability is driven by field and lab SOP details. Since almost any change in field or lab methods can change data values, the key question becomes: Are the method/protocol details similar enough to produce data truly comparable with other regional data sets that can provide perspective in data meaning?**

**Historically, data comparability has been controlled qualitatively. Newer quantitative approaches are recommended. Planners need to document that methods used will be optimal to help Parks answer identified question(s) as well as optimal for comparison with other important datasets and with important comparison criteria or standards.**

**Draft Departmental Information Quality Guidelines stress the need for a high degree of transparency for the data and methods to facilitate the reproducibility of such information by qualified third parties (http://www.mms.gov/whatsnew/PDFs/Info%20Quality%20Guidelines.pdf). The emphasis is on producing information is capable of being substantially reproduced, subject to an acceptable degree of imprecision (http://www.doi.gov/ocio/guidelines/515Guides.pdf).**

**It is impossible to reproduce data unless methods, SOPs, QA/QC and any other protocols are described in enough detail to allow others to reproduce exactly what was done.**

**Data comparability and interagency collaboration are also major areas of emphasis of the interagency groups trying to insure that data can be used for multiple purposes (see National Water Quality Monitoring Council summary at http://wi.water.usgs.gov/pmethods/mdcbfs.pdf).**

**Two key steps of determining data quality objectives are central to choosing and documenting protocols:**

**Determining methodological and practical technical specifications to achieve the information desired, and**

**Comparing different methods and choosing the one that meets the desired specifications within identified study limitations (Barbour, M.T., J. Gerritsen, B.D. Snyder, and J.B. Stribling. 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, Second Edition. EPA 841-B-99-002. U.S. Environmental Protection Agency; Office of Water; Washington, D.C., Section 4.1 (http://www.epa.gov/owow/monitoring/rbp/ch04main.html). By the time the planning group has finished with this step, decisions on maximizing data comparability while considering these two aspects of developing data quality objectives should be finalized.**

**Guidance documents helpful in choosing methods, SOPs, and more general protocols are listed as follows:**

## General Guidance Documents

**Default Clean Water Act methods, used by most States. These are detailed in 40 CFR Part 136.3 (Guidelines Establishing Test Procedures for the Analysis of Pollutants, Federal Register: August 18, 1998 (Volume 63, Number 159)] Environmental Protection Agency, see Web at http://www.access.gpo.gov/nara/cfr/cfrhtml_00/Title_40/40cfr136_00.html).**

**Most states use 136.3 methods for NPDES permit and general Clean Water Act (CWA) purposes and methods in 40 CFR part 141 for Drinking Water applications.**

**Many states have much more detailed protocols to be used in the field and lab. For example, the Texas protocols for both marine and freshwater are at http://www.tnrcc.state.tx.us/admin/topdoc/gi/252/swqmproc.pdf. This document contains not only some lab methods but also detailed field methods and QA/QC. Complete lab QA/QC is in a separate QAPP, not available on the Web. This QAPP calls for systematic error/bias (% recovery) performance of standards of 75-125%, and repeatability precision performance standards of 20-30% relative percent difference based on duplicate measures of the same substance under similar conditions. As is the case for many other states, many but not all Texas guidance**

**manuals and methods are on the Web (http://www.tnrcc.state.tx.us/water/quality/data/wqm/index.html#guidance).**

**On a National basis, a very helpful source of information on EPA and general methods is the beta version of the National Environmental Methods Index, an interagency product of the National Water Quality Monitoring Conference (NWQMC) is at www.nemi.gov. NEMI is a clearinghouse of environmental monitoring methods. The NEMI database contains method summaries of lab and field protocols for regulatory and non-regulatory water quality analyses. It is searchable over the World Wide Web. In NEMI, one can find the basic description of many methods in that data base, and on can also click on "links" on the home page, then click on EMMA, 4th link down on the page that comes up, to get the to the prototype version of the Environmental Monitoring and Measurement Advisor (EMMA), an interactive software (expert system) designed "to help you plan improved and cost-effective environmental monitoring projects" (http://infotrek.er.usgs.gov/doc/nemi/emma/). EMMA includes a recommended step-by-step decision process (expert system at http://infotrek.er.usgs.gov/doc/nemi/emma/EMMA-2N-Start.html.**

**Part C of this desk reference (detailed discussion of field methods and continuous monitoring probes, mostly but not totally related to freshwater) are on the Internet at http://science.nature.nps.gov/im/monitor/protocols/wqPartC.doc.**

**Methods for emerging issue compounds such as pharmaceuticals, hormones, and various organic wastewater contaminants (OWCs) have been summarized by USGS (K. K. Barnes, D. W. Kolpin, M.T. Meyer, E. M. Thurman, E. T. Furlong, S.D. Zaugg, and L. B. Barber. 2002. Water-Quality Data for Pharmaceuticals, Hormones, and Other Organic Wastewater Contaminants in U.S. Streams, 1999-2000, USGS Open-File Report 02-94, http://toxics.usgs.gov/pubs/OFR-02-94/#analytical). Quality assurance/quality control methods for these "emerging issue" compounds were also specified (http://toxics.usgs.gov/pubs/OFR-02-94/#quality).**

**For regulatory context monitoring related to municipal landfills or other RCRA work, SW-846 methods should be used (SW-846 online at http://www.epa.gov/epaoswer/hazwaste/test/main.htm).**

**For Superfund work, CERCLA contract lab program methods are given at http://www.epa.gov/superfund/programs/clp/ilm5.htm#sow and at links from that site (for example, methods for metals are at http://www.epa.gov/superfund/programs/clp/download/ilm/ilm52d.pdf).**

## Marine and Estuarine Methods

**EMAP protocols are used broadly by EPA and states in marine or estuarine environments and are suggested as a default choice for NPS monitoring in these environments. One major advantage is standardization up and down the coast lines and the development of a nationwide QA/QC plan for the program, that provides precision and systematic error performance standards for even simple measures like pH. If more frequent monitoring is needed to answer specific questions, it is suggested that more frequent monitoring be done but otherwise the EMAP**

**protocols be used to the extent possible to insure data comparability with EMAP data. The EMAP methods may be found in EPA. 2001. Environmental Monitoring and Assessment Program (EMAP): National Coastal Assessment Quality Assurance Project Plan 2001-2004. United States Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Gulf Ecology Division, Gulf Breeze, FL. EPA/620/R-01/002. (http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf. Although many States participate in the EMAP program, it is suggested that planners double check with the state involved to make sure they have no objection to the methods. If required by the State, specific State protocols and/or those in 40 CFR Part 136.3 (see Web at http://www.access.gpo.gov/nara/cfr/cfrhtml_00/Title_40/40cfr136_00.html) may need to be followed instead of or in addition to the EPA Coastal Assessment Documents.**

**Some states have protocols that generally consistent with those of EMAP but sometimes more detailed or rigorous. For example, the Texas field protocols for marine and estuarine areas (as well as freshwater) are at http://www.tnrcc.state.tx.us/admin/topdoc/gi/252/swqmproc.pdf.**

**Additional EPA guidance on bio-assessment and biocriteria can be found at http://www.epa.gov/ost/biocriteria/States/estuaries/estuaries1.html.**

## General Freshwater Methods

**For those desiring to standardize methods with comparable USGS freshwater data, see USGS guidance at http://water.usgs.gov/owq/FieldManual and/or USGS NAWQA protocols at http://water.usgs.gov/nawqa/protocols/doc_list.html (and/or http://water.usgs.gov/nawqa/protocols/methodprotocols.html) be used. Additional information on USGS lab methods and detection level information, and even real time data are at http://wwwnwql.cr.usgs.gov/, (general information) at http://toxics.usgs.gov/topics/measurements.html (measurements), and http://toxics.usgs.gov/bib/bib-methods.html (toxics).**

**Many states now have standard methods, SOPs, protocols and even QA/QC standards on the Internet. For example, see Wyoming SOPs at http://deq.state.wy.us/wqd/watershed/10574-doc.pdf, and Wyoming Quality Assurance Project Plan (QAPP) at http://deq.state.wy.us/wqd/watershed/10573-doc.pdf. Texas requires the use of field freshwater protocols at http://www.tnrcc.state.tx.us/admin/topdoc/gi/252/swqmproc.pdf**

**For aquatic biology studies, the Park Service will ordinarily use existing protocols to maximize comparability to other large regional datasets. In cases where States have not developed detailed biocriteria for use in water quality standards and have not developed detailed monitoring methods, general guidance on aquatic biological methods in various habitats may be found on the Internet at: http://www.epa.gov/owow/monitoring (choose biological assessment option). However, this is national EPA guidance, so monitoring planners should typically first need to check with the applicable State to see if State-specific protocols have**

**been developed. Many States have put their methods on the Internet. For example, Maryland's are at http://www.dnr.state.md.us/streams/mbss/mbss_pubs.html.**

**If the State has no detailed protocols related to biocriteria, NPS will typically use the guidance on the EPA website and/or USGS NAWQA, or other applicable regional protocols such as TVA's. Protocols selected should offer the added benefit of considerable comparable data within the region. Developing new protocols is to be avoided when possible, since if the Park is the only one using the protocols, fewer comparable, regional comparison/control sites are available and it is therefore typically impossible to tell if changes are due to human impacts such as aquatic pollution point sources or to regional climatic conditions (wet years, dry years, hot years, cold years, early snowmelt, late snowmelt, etc.) or to other regional factors such as air pollution.**

## Methods for Freshwater Lakes

**For general water quality studies, see EMAP guidance at John R. Baker, David V. Peck, and Donna W. Sutton (editors). 1997. Environmental Monitoring and Assessment Program Surface Waters: Field Operations Manual for Lakes. EPA/620/R-97/001. U.S. Environmental Protection Agency, Washington D.C. (http://www.epa.gov/emap/html/pubs/docs/groupdocs/surfwatr/field/97fopsman.html).**

**Guidance on biological monitoring of freshwater lakes, is available from EPA at http://www.epa.gov/owow/monitoring/tech/lakes.html.**

## Methods for Smaller Streams:

**For general water quality studies, see EMAP's 2001 draft guidance (Peck, D.V., J.M. Lazorchak, and D.J. Klemm (editors). Unpublished draft. Environmental Monitoring and Assessment Program -Surface Waters: Western Pilot Study Field Operations Manual for Wadeable Streams. EPA. U.S. Environmental Protection Agency, Washington, D.C., http://www.epa.gov/emap/html/pubs/docs/groupdocs/surfwatr/field/fomws.html.**

**Guidance on biological monitoring of wade-able streams, is available from EPA at http://www.epa.gov/owow/monitoring/rbp/.**

## Methods for Larger Streams and Rivers:

**For non-wade-able streams, see EPA 2000. Environmental Monitoring and Assessment Program-Surface Waters: Field Operations and Methods for Measuring the Ecological Condition of Non-Wadeable Rivers and Streams (http://www.epa.gov/emap/html/pubs/docs/groupdocs/surfwatr/field/Intro_mat.pdf). Background information on fish health and contaminants in fish of several large rivers may be found at http://www.cerc.usgs.gov/data/best/search/.**

## Guidance for Wetlands:

**The 40 CFR guidance of EPA and other general and freshwater guidance documents listed above should be helpful for wetlands, and bio-assessment information is available at http://www.epa.gov/owow/wetlands/bawwg/.**

## Typical NPS qualitative data quality objectives:

**The methods, standard operating procedures (SOPs), and any other more general field and/or lab protocols selected shall be optimal to help Parks answer identified question(s), and shall be documented in the plan in enough detail so that they could be reproduced by third parties. This is required not only for data comparability but also to meet a NPS "good science" data quality objective that all data produced should be not only repeatable by the same investigator and lab, but also reproducible by independent parties (see definitions of these two terms in the definitions section at the end of the appendices). Detailed metadata on all field and lab methods must be documented so that others could reproduce the study.**

**To the extent possible, Standard Operating Procedures (SOPs), methods, and more general protocols selected should be those that insure data comparability in general and that the data are comparable with other large regional data sets and are acceptable to interested regulatory groups. In other words, in concert with the recent push to make water monitoring data useful for multiple purposes, when practicable, protocols/SOPS used should be selected that ensure that the data collected should be useful for both regulatory purposes and for general monitoring of status and trends.**

**Therefore, networks are generally encouraged to use standardized protocols already developed and in wide use in the region rather than inventing new ones when it is not necessary to do so. In the case of aquatic sampling, these might include State protocols, USGS WRD protocols, USGS NAWQA protocols published in the USGS Field Manual, NOAA protocols, EPA EMAP protocols, EPA and State Bioassessment Protocols, and other standard protocols such as those listed in the sections above. The networks are not encouraged to change these types of well standardized protocols, but merely to augment the standardized protocol with some material in the narrative or SOPs to better explain how they implement the protocol in their particular study areas. For example, an SOP might describe how to locate and travel to each of the sampling sites, or another SOP might describe where the equipment is stored, and how many copies of the field form to copy before going out into the field. They may also need to add missing sections such as personnel training or what it costs to implement the standardized protocol. Networks are told to keep an archive of all the versions of the sampling protocol used over time, and to put a field in the database that references which version of the protocol they used (Steven Fancy, NPS, I&M Program, Personal Communication, 2003).**

**Before monitoring begins, if one already has considerable precision reproducibility data from two different labs or two different methods, two different operators (or anything else that has changed in the measurement process) one should compare them quantitatively to determine if they are comparable enough to meet data quality objectives specified for precision in sections V-D and V-E.**

**After the monitoring has proceeded for a few years, the performance observed in meeting data quality indicator goals for comparability to both regulatory and general regional status and trends data sets should be quantified. For details on how to make both qualitative and QUANTITATIVE comparability decisions on both new and older data, see section 3.4 of latest proposed EPA guidance [EPA 2001. Guidance on Data Quality Indicators (EPA QA/G-5i) at http://www.epa.gov/quality/.**

**If the goal is to make sure that data is comparable to regional USGS data, or other regional status and trends data sets, the percent of the data that is expected to meet this goal should be specified in the plan.**

**If the monitoring is done for regulatory purposes, SOPs, detailed methods, and QA/QC should in all cases meet the requirements of the regulatory agency.**

**In concert with generic Vital Signs Guidance, the plan should document the methods by "Giving an overview of each sampling protocol that will be used to monitor the vital signs. The full protocols should be included in an appendix. The overview should summarize the material in the protocol, including a narrative for each protocol, an overview of the resource issue being addressed, specific measurable objectives, sampling design, field methodology, data analysis and reporting, personnel requirements, and operational requirements. Generally accepted Standard Operating Procedures for the collection of data for constituents that may serve as water quality vital signs are provided in the Vital Signs Desk Reference" (http://science.nature.nps.gov/im/monitor/deskref.htm).**

**After the monitoring has proceeded for a few years, the performance observed in meeting data quality indicator goals for comparability to both regulatory and general regional status and trends data sets should be quantified. For details on how to make both qualitative and QUANTITATIVE comparability decisions on both new and older data, see section 3.4 of latest proposed EPA guidance [EPA 2001. Guidance on Data Quality Indicators (EPA QA/G-5i) at http://www.epa.gov/quality/.**

**The concept of protocols is more general than methods or SOPs and accordingly should include a detailed discussion of Quality Assurance/Quality Control measures used to insure that data collected will be considered credible by those who will be using it, including NPS managers, state agencies, and other federal agencies. These steps include all of the QA and QC steps listed herein.**

**SOPs should contain details of sampling methods, sample handling, sample custody, analytical methods, instrument/equipment testing and inspection, instrument maintenance and calibration, calibration frequency, and methods to be used for accepting and using supplies and consumables that come in contact with chemical samples. As detailed elsewhere herein, to the extent appropriate to facilitate data interpretation, metadata details on methods and SOPs shall also be reported in STORET.**

**Requirements for maximum holding times (sample storage time limitations), preservation methods required, the use and storage of chain of custody forms, and approved containers should be listed in the plan. Unless otherwise justified in the plan both status and trend monitoring and regulatory monitoring should utilize standard EPA protocols for holding times, sample containers, and preservation (as summarized in 40 CFR Part 136.3, see Web at http://www.access.gpo.gov/nara/cfr/cfrhtml_00/Title_40/40cfr136_00.html. Most states use 136.3 recommendations for NPDES permit and general Clean Water Act (CWA) purposes. The exception for the CWA relates to Drinking Water Act applications (see 40 CFR part 141 for Drinking Water Methods).**

**When no regulatory issues are involved, and/or the State has no objection, it is suggested that either EPA EMAP or USGS protocols be used for default guidance, depending on the state's needs and who has the preponderance of data with which the new NPS data will be compared**

**See appendix V-A.1 for details.**

### V-A.2 Selection of a Chemical Lab (if applicable).

**Some water quality parameters, such as pH, dissolved oxygen, temperature, PAR, and conductivity derived measures (specific conductance and salinity), are typically best measured in the field as the samples are collected. For some other parameters, field measurements are not practical or do not produce the best results so chemical laboratories are typically used.**

**One of the biggest mistakes a monitoring team can make is to choose a lab that has not passed difficult and independent State and federal (Federal National Environmental Laboratory Accreditation Program or NELAP, see http://www.epa.gov/ttn/nelac/accreditlabs.html) accreditation/approval QA/QC checks, optimally including blind-sample round-robin trial analyses of proficiency test (PT) standards to see if the answer the lab provides is close enough to known (certified correct) ranges to pass QC performance standards.**

**Rigorous and independent checks are needed to insure that the lab chosen can produce accurate (low uncertainty in the value obtained versus the true value) and comparable (with other major data sets of interest) results. Federal and state agencies that have run round-robin testing programs have determined that many candidate labs cannot pass such checks.**

**The Park Service does not have its own laboratory accreditation or approval program. Therefore, in picking labs that can pass rigorous tests, it is probably safest to select labs that have gained approval of at least one (two is even better) federal agency that runs programs to approve laboratories after rigorous testing related to the environmental media of interest. The best of these programs typically require not only paper checks to ensure that QA/QC programs are in place, but also require candidate labs to demonstrate adequate performance on analyses of blind samples in an inter-laboratory round-robin analyses of an NIST-certified or other high quality reference materials.**

**The Fish and Wildlife Service (FWS) runs such a program (for a list of FWS approved contract laboratories, see http://www.pwrc.usgs.gov/pacf/, then click on analytical laboratories). These labs will usually give the NPS Fish and Wildlife Service prices and QA/QC specifications). The Fish and Wildlife Service (FWS) is a sister agency to the National Park Service within the same directorate within the Department of Interior. The FWS also produces considerable contaminants data at Wildlife Refuges, and using FWS approved contract labs and QA/QC specifications is one logical way to ensure data comparability with these other Department of Interior datasets.**

**NOAA's Status and Trends (Mussel Watch) program has a performance-based (PBMS) lab cross comparison program run by NIST that has been held out by an interagency group as a model example of how to do round robin comparisons on reference samples reflecting environmentally relevant concentrations and matrices (J. Diamond et al. 2001. Towards a definition of performance-based laboratory methods. A position paper of the National Water Quality Monitoring Council Methods and Comparability Board, Technical Report 01-02, Web: http://wi.water.usgs.gov/pmethods/PBMS/nwqmc.0102.pdf). Other labs may join the NIST/NOAA round robin comparisons for a fee.**

**Analyses of contaminants in solids such as sediments or tissues can be even more difficult than similar analyses in water. Those in EPA who provide advice on fish tissue health advisories recommend that labs contracted should be involved in round-robin QA programs, such as that administered by NOAA in conjunction with its National Status and Trends (NS&T) Program (http://www.epa.gov/ost/fishadvice/volume1/v1ch8.pdf).**

**EPA's estuarine/marine EMAP program agrees with the modern consensus that performance based (PBMS) lab selection is good, but should include round robin inter-lab comparisons (for more information see appendix V-A.2 and EPA recommendations at http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf.**

**The United States Geological Survey (USGS) has its own outside lab approval program, and the Department of Defense (DOD) has run similar checks of its own.**

**For more information on the programs listed above and recommendations of the Methods and Data Comparability Board (MDCB) Accreditation Workgroup of the federal (interagency) National Water Quality Monitoring Council, see appendix V-A.2.**

**Typical NPS Data Quality Objectives:**

**For laboratory analyses of water, tissue, soil, or sediment samples, a first choice to ensure data comparability and quality would be to utilize labs that have passed multiple federal round-robin laboratory inter-comparison checks. For example, the Texas A. and M. GERG laboratory (mentioned as merely as an example, there may be others) has passed recent QA/QC round robin checks of both and the Fish and Wildlife Service and the NOAA inter-comparison program run by NIST.**

**Second choice, but still acceptable would be to utilize any lab that passes round-robin, blind-sample QA/QC checks of at least one Federal agency (USGS, NOAA, NIST, U.S. Army, etc.). Optimally, the lab should thereby also pass NELAC proficiency tests with blind proficiency test (PT) standards (http://www.epa.gov/ttn/nelac/standard/chapter2.pdf). An example of a federal approval program that includes a major Federal National Environmental Laboratory Accreditation Program (NELAP) component as well as a federal approval process is the one run by the U.S. Army. The Army's USACE evaluates the PT results that laboratories submit to the NELAP state accrediting authorities and performs on-site inspections on the basis of NELAP and supplementary USACE programmatic QA/QC requirements (Thomas Georgian, U.S. Army, Personal Communication, 2003, for more details see http://www.environmental.usace.army.mil/info/technical/chem/chemval/chemval.html).**

**Third choice (and absolute minimum) would be to choose any lab that appears on national NELAP-approved labs lists for the regulatory program of interest (Clean Water Act, Drinking Water Act, CERCLA etc., see http://www.epa.gov/ttn/nelac/accreditlabs.html), One goal of NELAP is to make sure that laboratories have an adequate QA/QC system conforming to ISO 17025, or equivalent. For monitoring being done for other specific regulatory issues (such as those related to water quality standards impairments, drinking water issues, RCRA, or CERCLA for example), the chemical labs selected should also be acceptable to the regulatory agency administering that program. Again, the place to start with this assessment is the NELAP list, which lists approval for use in various regulatory programs.**

**In picking a laboratory, planners should also consider the judgment of the State where the samples are collected. For water samples, if the State is one agency that will be using the data for regulatory purposes, such as TMDL or water quality standards exceedances (or otherwise needs to be convinced about the quality of the data in any way), it is also important to choose a lab that is acceptable to the State agency. This often means picking a lab that has passed State accreditation checks or can produce a detailed Quality Assurance Project Plan (QAPP) that is considered acceptable to the State. Decide who will have a regulatory interest in the data. If data is to be used primarily or only for state regulatory purposes, it is especially critical that**

**labs have been approved not only by NELAP or other federal programs, but also by the State regulatory agency or its accreditation affiliate. In most cases, it should be sufficient to confirm that the NELAP accrediting agency documented in the NELAP list (see paragraph above) is a State agency, but when in doubt, ask the State Agency that may later be using the data for regulatory purposes.**

**To the extent possible, labs chosen should also be those that are able to achieve QUANTITATIVE detection limits below applicable water quality standards or other key data comparison benchmarks (see listing of data comparison benchmarks useful as thresholds or decision rule trigger points in section IIII-B on Decision Rules).**

> **Note: Terminology has been quite varied by agency and lab, but whether they are called minimum levels (MLs) or practical quantitation limits (PQLs), minimum reporting levels (MRLs) or the limits of quantitation (LOQs), minimum quantitation limits (MQLs), or laboratory reporting levels (LRLs), or something else, it should be made clear that when possible, it is important to have QUANTITATIVE detection limits (not just semi-quantitative Method Detection Limits/MDLs) below key comparison benchmarks and decision rule threshold levels. Numerous EPA methods state "The MRL should be established at an analyte concentration either greater than three times the MDL or at a concentration which would yield a response greater than a signal to noise ratio of five (For example, see (EPA, 1999. Method 314.0, Determination Of Perchlorate In Drinking Water Using Ion Chromatography at http://www.epa.gov/safewater/methods/met314.pdf).**
>
> **Another way of saying this is that, to the extent possible, labs chosen should be those that are able to achieve QUALITATIVE Detection Limit (usually a Method Detection Limit (MDL or LT-MDL) far enough (typically 2-10 times) below the water quality standard or other comparison benchmark or threshold level to be considered quantitative by the analytical laboratory. In other words, laboratory limits of quantification are typically established by the lowest instrument calibration standard (or are set at higher concentrations) and these quantitative detection limits are typically greater than semi-quantitative detection limits (such as MDLs) that determine only presence-absence of an analyte**.

**Likewise, the NPS should only use laboratories that can pass quantitative acceptance standards (QC performance standards) for other data quality indicators (precision, systematic error/bias, accuracy, and comparability) at least as stringent as required by measurement quality objectives listed in the**

detailed study plan/QAPP. For details on precision and systematic error, see sections V-E and V-F.

### V-A.3 Selection of a Project Leader and Monitoring Staff

Chapter 8 of the generic VS guidance (Outline for Vital Signs Monitoring Plans, 2003, http://science.nature.nps.gov/im/monitor/docs/monplan.doc) states "For field sampling efforts to be performed in house, describe how they will be supported in terms of staff training and/or previous experience, field equipment to be dedicated to the effort (vehicles, instruments), anticipated in-house lab work to support operation, maintenance, and calibration of equipment and its documentation, and the necessary safety considerations in performing field tasks" (Outline for Vital Signs Monitoring Plans, op cit.). So some of the details of "who will do the work" will be covered in Chapter 8. However, at least a brief summary discussion of how choices made in "who will do the work" have been made in a way that maximizes data comparability and quality should be given in chapter V (Protocols) of the Monitoring Plan.

Rationale: Who will oversee and do the work and recurrent training can impact the SOPs used as well as data comparability and data quality. These are all Quality Assurance basics. Therefore, some overview discussion of how the staff selected for the work and how recurrent training will be accomplished to help optimize data comparability should be presented in Section V.

Each protocol narrative should provide more details on "personnel requirements, training procedures, and operational requirements (K.L. Oakley, L.P. Thomas, and S.G. Fancy, 2003. Guidelines for long-term monitoring protocols. Wildlife Society Bullet 31(4), reprinted on Internet at http://science.nature.nps.gov/im/monitor/protocols/ProtocolGuidelines.doc).

The plan should at least briefly summarize how the project leader and staff were chosen to help assure data comparability. Individual protocols (especially the QC protocol) shall document that the project leader and monitoring staff selected are not only proficient to do the work but are also willing and able to ensure data comparability and QA/QC sufficient to satisfy data quality objectives outlined in the plan and in QC SOPs attached to each protocol, as well credible data statutes of State regulatory agencies and QA/QC documentation of federal regulatory programs such as the various programs in EPA. Most of these require that a quality assurance project plan (QAPP) be included in detailed study plans.

## V-B. Sensitivity, Detection Limits, Resolution, and Calibration

Measurement sensitivity is a basic QC data quality indicator (DQI) that has historically been addressed by controlling and documenting detection limits.

**Chemical lab detection limits (a special case of sensitivity at the lowest end of the quantitative measurement range), are not applicable to certain parameters, such as field measured pH, so alternative ways control measurement sensitivity are recommended below.**

**Generic Vital Signs guidance states that "Statistical detection limits, given typical sample variability and chosen sample sizes, shall be low enough to insure that such threshold values or trigger points can be detected" (Outline for Vital Signs Monitoring Plans, 2003, http://science.nature.nps.gov/im/monitor/docs/monplan.doc). The types of sensitivity discussed herein include:**

1. **Measurement sensitivity, pertaining even to a single value or data point (the topic of this section).**
2. **Statistical/study design sensitivity to detect a change (Section V-B.7),**
3. **The sensitivity of values to be protected to changes in model components, including potential measures (Potential Vital Signs, section II-C). This is a broad type of relationship or model sensitivity.**

**There is a strong relationship between the concepts of measurement sensitivity, detection limits, calibration, measurement resolution, and measurement uncertainty. The terminology and usage of these words and phrases in environmental monitoring has been varied and often confusing or conflicting. The widespread confusion on terminology could be resolved if everyone spoke same language, preferably by using NIST/ISO terminology already standardized for U.S. and world scientific and engineering communities. However, in spite of widespread confusion, the following comparison basics seem relatively clear:**

**Sensitivity involves the ability measure with a defined confidence. The (particular type of) sensitivity indicators of primary interest to EPA related to water quality and contaminants work are detection limits. (http://on-linelearning.ca/idec4433/epaqaqc2000/g5i-prd.pdf).**

**Signal to noise ratios (S/N) are one of the better ways to determine sensitivity, but there are no generally standardized ways to calculate S/N and often practical considerations result in determining sensitivity from replicate precision measuring of single low (or blank) level samples. The more sensitive a method is, the better able it is to detect lower concentrations of a variable," and the lower the detection limits**

**A minimum detection limit is special kind of sensitivity, typically at the lowest end of the range of values in which quantitative measurement is possible. Detection limits (rather than resolution) relate to the capability of a method or instrument to correctly discriminate between small differences in analyte concentration at very low concentrations (or other signal values) (http://www.epa.gov/volunteer/qapp/qappappa.pdf).**

**Calibration includes efforts to insure an instrument is measuring with an acceptably low amount of systematic error/bias.**

**Measurement uncertainty is the quantitative uncertainty in a single measurement result. In optimal measurement ranges (not below quantitative detection limits), it might be argued that measurement uncertainty is one of the**

**better ways to estimate sensitivity in general. However, standard NIST/ISO methods (N. Taylor and C. E. Kuyatt. 1994. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results NIST Publication TN 1297, http://physics.nist.gov/Document/tn1297.pdf) to calculate measurement-uncertainty are not applicable for very low value (below quantitative detection limits) ranges of measurement. In these regions, detection level explanations of quantitative and qualitative uncertainty take over.**

**Terminology used for electronic meters and probes used in field measurements has often been non-standard and confusing. Neither sensitivity nor detection limit specifications are typically provided by manufacturers. It is often not clear what is meant by advertised accuracy or resolution specifications. Neither can automatically be trusted to reflect real-world sensitivity, detection limits, or measurement uncertainty. Thus, for the applicable quantitative measurement ranges, it is typically necessary for users to determine NIST/ISO expanded measurement uncertainty and/or more general estimates for measurement sensitivity (such as multiples of the sample standard deviation of the lowest measurable concentration, see more details below) in the particular field environmental conditions encountered.**

**To avoid confusion, the word resolution should not be used as a synonym for sensitivity or for related concepts like detection limits, precision, or measurement uncertainty. The term should not be used at all in study plans/QAPPs unless is clearly defined and the difference between resolution and the other terms listed above are made clear.**

**Each of these topics are discussed in more detail below, and in greater detail in appendix V-B.**

## V-B.1 Sensitivity in General

**NIST has no definition for sensitivity. However, a publication recommended by NIST and published by ISO in 1994, the International Vocabulary of Basic and General Terms in Metrology; ISO/TAG 4 1994 defines sensitivity as a "change in the response of a measuring instrument divided by the corresponding change in the stimulus" and notes that "the sensitivity may depend on the value of the stimulus." (http://www.measurementuncertainty.org/mu/glossary/index.html). In this context, perfect theoretical sensitivity would be a ratio of 1/1=1.**

**However, in environmental study plans and QAPPs, we don't typically see measurement sensitivity expressed as ratios, fractions, correlation coefficients, or the like. The exact true value of the stimulus is sometimes hard to quantify. More importantly, the type of measurement sensitivity we usually care about the most is the smallest change in a measurement or observation that accurately (with low uncertainty) reflects a true change in the "signal," the true magnitude of the environmental parameter of concern. This type of measurement sensitivity is usually expressed as a detection limit at the lowest end of practical quantitative measurement range. Sensitivity in optimal quantitative measurement ranges is better expressed in alternative ways explained in sections V-B.2.**

**In this document, sensitivity in general is discussed for three levels of organization or concern:**

4. **Measurement sensitivity pertaining to a single data point is discussed in section V-B.2 (just below). Related sensitivity issues that relate to measurement sensitivity or the level of organization of a single data point issues are discussed in sections VI.B.3-6.**
5. **Statistical/study design sensitivity, pertaining to multiple data points and the ability of the statistical design to detect statistical changes considered important, is discussed in the section V-B.7.**
6. **The sensitivity of a proposed vital sign to some other broader value or concept (for example, values to be protected, desired future condition or aquatic ecological integrity was discussed earlier in section III-C.**

### V-B.2 Measurement Sensitivity

**In Quality Assurance Project Plans (QAPPs), one usually sees estimates of measurement sensitivity expressed as detection limits, optimally both quantitative and semi-quantitative detection limits. Detection limits constitute a particular kind of measurement sensitivity applicable to values at the lowest end of the measurement range where quantitative measurements are possible.**

**Although sensitivity is listed by EPA as data quality indicator QC basic to be included in QAPPs, detection limits have so often been equated with sensitivity that EPA's new STORET data base does not even have a designated place (other than general comment fields) to report any kind of sensitivity other than detection limits.**

**Detection limits are particularly applicable to toxic chemicals that may have effects at very low levels. Detection limits are not only the most common way to control sensitivity controlled in the past, they are still the first choice for controlling sensitivity in the measurement of concentrations or other signals at the lowest end of practical quantitative measurement.**

**However, detection limits are not designed to determine sensitivity as a change "we really believe" in the normal quantitative measurement range (optimal measurement ranges well above detection limits). Measurement uncertainty (section IV-F.2) is a logical way to do this. Some parameters are always present in the environment at values well above detection limits, so detection limits are less important in these cases. Detection limits simply cannot be calculated for parameters such as pH and so measurement sensitivity must be controlled in some other way. Measurement uncertainty can be calculated from the results of two other data quality indicators commonly controlled in QAPPs, measurement precision (see sections V-D and E) and systematic error/bias(see sections V-F and G).**

**In cases where all measured values are well above detection limits, or in cases where detection limit methods are simply not applicable, sensitivity in the quantitative measurement range needs to be controlled and documented in terms of measurement uncertainty (see Section IV-F.2). When possible it is optimal to control low level sensitivity in terms of detection limits and higher level sensitivity in terms of measurement uncertainty.**

> Note: In the past, measurement uncertainty in the quantitative measurement range has too often been largely or totally ignored, almost as though the investigators were assuming that each measurement or observation was perfect. No measurement or observation is perfect, each is an estimate. Just as confidence intervals express the uncertainty about a mean, there is an interval of uncertainty around each measurement data point, One way to express measurement sensitivity for values above detection limits to calculate measurement uncertainty for each data point.
>
> Some biological or ecological measurements are linear (number of taxa, number of individuals, etc). To estimate systematic error, an expert's counts can be considered "right" or "expected" and a trainee's counts can be considered wrong (or at least less than optimal). In these cases, measurement sensitivity can often be estimated as measurement uncertainty in the normal ways. In biology and ecology, measurement uncertainty and sensitivity need to be controlled, just as in chemical or physical measurements

**There are some cases where neither standard detection limits nor measurement-uncertainty are applicable or are easily calculated. In these cases, alternative ways to estimate sensitivity based on multiples of the sample standard deviation for precision (see details farther below), may be reported. This is typically a less than ideal and sometimes less than a complete way to bound the smallest measured change that "we really believe" than the two options (detection limits and measurement uncertainty) discussed above. However, sometimes this is the only option easily or practically accomplished, and controlling measurement sensitivity in this way is typically better than not controlling sensitivity at all.**

**Therefore, for analytes for which standard EPA or USGS defined MDL-related detection limits (see section below on detection limits) cannot be easily calculated, it is suggested that sensitivity should be estimated as a limit of quantification (LOQ) in the easiest of the following optional ways:**

> "The LOQ can be defined in a number of ways, such as the background response plus ten times the standard deviation of the lowest measurable concentration, ten times the signal-to-noise ratio of the baseline noise, ten times the standard deviation of the lowest measurable concentration, etc." (EPA. 1998. Final report of the FIFRA Scientific Advisory Panel open meeting held in Arlington, Virginia, on March 24-25, 1998, http://www.epa.gov/oscpmont/sap/1998/march/chapb-1.pdf).

**In many situations where all else fails, ten times the standard deviation of the lowest measurable concentration (or other "signal"), one of the options provided above, can usually be used to estimate measurement sensitivity as an LOQ. This would work even for field determinations of pH, as long as one accepted a readily available calibration standard (for example, one for a pH of four) as a readily obtainable and reasonably relevant to environmental levels, practical "lowest**

**measurable concentration." This could usually also be done for biological, ecological, and habitat measurements or observations, if one accepts the lower end of the range of observations as the "lowest measurable concentration" (or signal).**

## SIGNAL TO NOISE (S/N) RATIOS

**Signal to noise (S/N) ratios are frequently key concepts not only in detection limits but also in more general concepts of measurement sensitivity. To control sensitivity in a the measurement process, typically we not only want to estimate signal (such as a true concentration of a chemical or the true magnitude of light energy in PAR), but if we report a change in that value, we want to have some confidence that our reported change really reflects a true change in the strength of the parameter being measured (the "signal"), rather "noise." In this context we define noise as "random up and down fluctuations due to imperfections (such as interferences) in the measurement process." Noise is typically estimated from at least seven replicate precision repeatability samples at lowest measurable concentrations or in blanks.**

**S/N ratios of 2 to 5 are typically associated with semi-quantitative detection limits such as limits of detection (LODs) or method detection limits (EPA's MDLs or USGS LT-MDLs). S/N ratios of a minimum of 5 to 10 are typically associated with quantitative detection limits such as limits of quantitation (LOQs), EPA's PQLs, or USGS MRLs (see appendix V-B.4 for more details.**

**If signals remain constant, high signal-to-noise ratios indicate not only good sensitivity but also high precision as measurement repeatability. If noise remains constant, stronger signals result in higher signal to noise ratios and better measurement sensitivity. However, at the lower end of the quantitative measurement scales, signals tend to become weak in relationship to noise, which is why methods to estimate detection limits are needed.**

**Although S/N ratios are appealing and relevant, there are no universally accepted definitions of how to calculate S/N ratios, and they are not applicable to parameters such as biochemical oxygen demand (BOD), total suspended solids (TSS), fecal coliforms, and pH. There are numerous potential difficulties and complications with S/N ratios (see appendix V-B.2 for details).**

**To avoid confusion with a similar concept at a higher level of organization (statistics summarizing true variability of a parameter in the environment), the word "noise" should be used only in the context of measurement imprecision (random up and down fluctuations due to imperfections in the measurement process). True variability in parameters in the environment should be called something else, such as true variability or true heterogeneity to distinguish these concepts from measurement noise**

**In the biological/ecological realm, only a handful of studies have measured the relationship between the question to be answered and the ability of different levels of taxonomic resolution to distinguish sensitivity patterns (e.g., differences between reference and impacted sites). Will Clements was able to distinguish a strong signal of metal pollution across a broad geographic region by looking at abundance of one group of benthic invertebrates, Ephemeroptera (mayflies).**

**However, species level identification (a finer scale of resolution) was necessary to characterize this variation within a single watershed. So different levels of taxonomic resolution (fineness of scale) may be necessary to obtain the sensitivity needed to answer different questions (Will Clements, Personal Communication, 2003, for more information see: Clements, W. and M. C. Newman, 2002. Community Ecotoxicology, John Wiley & Sons, 350 pp., http://www.wiley-vch.de/publish/dt/books/bySubjectNU00/bySubSubjectNU/0-471-49519-0/?sID=d05b).**

**One can avoid some confusion by not using the phrase "measurement errors." Instead use more standard phrases and terms as those discussed herein, such as measurement uncertainty, measurement sensitivity, and systematic error/bias. Statements such as "errors in the analytical measurements should be no greater than the natural variability of the parameters of interest" are somewhat akin (on a different scale of organization) to saying a S/N would be OK. Such statements should not be accepted for typical environmental monitoring scenarios. Of course the total measurement uncertainty should be less than the natural variability but for typical monitoring of chemical or physical parameters in the environment, the signal should typically be 2 to 10 times smaller than the natural variability. This is due not only to S/N concepts, but also because we typically want to understand variability characteristics in time and space, so typically errors in analytical measurements need to be MUCH smaller than the range or even the sample standard deviation of the natural variability of the parameters of interest. Another way to think of this is that just because a natural process has a very high degree of variability, is not necessarily an excuse to choose an extremely crude measuring system. Doing so might sometimes open the door to non-neutral parties trying to use crude methods as one way to find no differences between impacted and more pristine sites. A simple example may help illustrate this:**

> **If temperature ranged from 0 to 100 degrees Fahrenheit, and one wanted to understand how the temperature varies in time in space, one would clearly not choose a thermometer that measures to the nearest 100 degrees/**

**Furthermore, part of the problem with the statement that "errors in the analytical measurements should be no greater than the natural variability of the parameters of interest" is that it jumps from one level of concern/organization (measurement sensitivity and uncertainty on each data point) to another (statistical study design sensitivity and uncertainty related to statistics that summarize many data points). Therefore, it would be questionable whether even the traditional measurement level signal to noise ratio (S/N) of 10 or greater often associated with quantitative measurement sensitivity would be sufficient considering two levels of organization/concern are being addressed. It would be more relevant if both of the values being compared were on the same level of organization/concern.**

**For more detailed information on measurement sensitivity in general, and a more detailed discussion of the many complexities of S/N ratios, see appendix V-B.**

# DQOs for Measurement Sensitivity in Field Estimates of Biological or Physical Habitat Condition:

**Some states have specialized bio-assessment methods to address sensitivity or detection limits. For example, Wyoming Bioassessment guidance for a "method detection limit" involves determining if a standard and consistent level of identification can be maintained for organism-spiked samples (for more information see http://deq.state.wy.us/wqd/watershed/10574-doc.pdf).**

**However, most states and biological protocols have no such specialized methods, so more generic methods (such as those detailed below) need to be used.**

**When practical, measurement sensitivity should be estimated by calculating expanded measurement uncertainty as explained in section IV-F.2. This is the preferred option for estimating quantitative measurement sensitivity for parameters for which one cannot easily calculate normal detection limits. Measurement uncertainty is the estimate that best defines the smallness of a change "we really believe" is quantitatively valid.**

**In cases where it is difficult or impossible to estimate measurement uncertainty, or estimates of sensitivity based on S/N ratios or normal detection limit concepts (such as those used the laboratory measurement of chemicals), it is recommended that sensitivity be estimated from precision in one of two ways recommended by EPA (EPA. 1998. Final report of the FIFRA Scientific Advisory Panel open meeting held in Arlington, Virginia, on March 24-25, 1998, http://www.epa.gov/oscpmont/sap/1998/march/chapb-1.pdf):**

1) **SEMI-QUANTITATIVE (presence/absence) measurement sensitivity may be estimated as a limit of detection (LOD). The LOD is the point at which "a measured value becomes...larger than the uncertainty associated with it." The LOD is calculated as three times the (sample) standard deviation (of at least 7 replicate precision samples) at the lowest measurable concentration (or lowest signal that can be practically obtained).**
2) **A QUANTITATIVE limit of quantification (LOQ) may be calculated as ten times the (sample) standard deviation (of at least 7 replicate precision samples) at the lowest measurable concentration (or lowest signal can be practically obtained.**

**For biological or ecological estimates, "the lowest signal that can be practically obtained" is most easily obtained as "the signal" of a typical sample at a control (pristine) site. In other words, don't agonize at finding a sample with the lowest possible signal, just use a typical single sample from a pristine/un-impacted (as possible) control site. Keep in mind that the standard deviation is based on repeated measures of a single sample, not multiple samples. Therefore, the standard deviation**

**is estimating variability in the measuring system rather than heterogeneity of multiple samples. In this context, the standard deviation is an estimate of "noise" in the sense it is estimating measurement imprecision (random up and down fluctuations due to imperfections in the measurement process). Although basing sensitivity solely on multiples of the standard deviation of measurement noise (imprecision) is not ideal, it is based on common practice and is easily done. Sensitivity estimates obtained in this way are perhaps not as sophisticated as detection limit concepts that call for 99% confidence, but can be justified as a rough (better than nothing) estimate of measurement sensitivity.**

**The plan should document that the proposed measurement sensitivity will be sufficient to answer specified questions in light of other limitations on sensitivity at a higher levels (such as study design/statistical sensitivity as discussed below in section V-B.7).**

### V-B.3 Measurement Sensitivity of Electronic Meters and Probes Used in Field Measurements

**Monitoring planners should be aware that what field-measurement probe manufacturers advertise as resolution "specifications" are typically not calculated in consistent ways, nor are they necessarily good real-world (field condition) estimates of:**

1. **Measurement sensitivity, or**
2. **Uncertainty in low signal-level sensitivity (detection limits).**

**Some manufacturers seem to consider either precision as reproducibility (multiple meters or probes) or rounding rule (only the last digit displayed has any uncertainty) factors into their development of resolution specifications. Some say that their advertised resolution specifications may be used as a "starting place" rough estimate of ideal (lab) condition semi-quantitative measurement sensitivity.**

**However, other manufacturers just say that resolution simply reflects the number of digits or decimals displayed by the meter, without regard to confidence. This is probably the safest way to consider resolution and not to confuse the word with measurement sensitivity or detection limits.**

**Since the word resolution is not typically used in quality assurance project plans, it is probably better to not use the word at all in the context to monitoring water quality or contaminants, but to instead use the more conventional and universally defined and understood phrases such as "measurement sensitivity" and "detection limits."**

**The market place is understandably competitive, and in candid off-the-record moments, staff of more than one manufacturer have admitted to suspecting that "specsmanship" (they even have a word for the concept) is being used by their competitors to exaggerate the fineness of scale of resolution specifications and thus gain competitive advantage in marketing. Field trials have confirmed that manufacturers specifications for resolution do not reflect "what we really believe"**

**in long term field deployments of continuous monitoring sensors (Remote Underwater Sampling System/RUSS Quality Assurance Summary, University of Minnesota Duluth, Natural Resources Research Institute Water Educational Website at http://wow.nrri.umn.edu/wow/under/qaqc.html, see also http://lakeaccess.org/QAQC.html):**

**Therefore, Park Service users of meters and probes used to measure in the field must not simply use resolution specifications as estimates of measurement sensitivity and must instead estimate actual measurement sensitivity performance IN THE FIELD. Sensitivity performance will typically vary depending on many real-world factors such as the harshness of the measuring environment, how often the meter is calibrated and what has happened to it since that calibration, the expertise of the person doing the calibration and measurements, and many other factors.**

**Manufacturers typically provide a range for which the instrument meets accuracy and resolution specifications. Specific "detection limits" are typically not specified by manufacturers of field measurement probes and meters, and are sometimes not applicable. Therefore, users are presumably particularly unsure of the quantitative uncertainty in accuracy of readings below the applicable quantitative measurement concentration ranges specified by the manufacturer.**

**It is very difficult or even impossible to develop sensitivity estimates akin to detection limits for some parameters. For example, there are always hydrogen ions it is impossible to come up with a water-based blank for pH, so typical EPA 40CFR part 136 methods for determining semi-quantitative detection limits do not work well. Examples parameters, that don't lend themselves to easy estimates of sensitivity, as detection limits include (list is probably not complete):**

1. **PAR (light attenuation in lakes or marine environments)**
2. **Bacterial Counts (fecal coliforms, E. coli, etc.)**
3. **Taxonomic Identification of Very Small Invertebrates or Other Difficult Taxa**
4. **Judgment Habitat Observations (Percent Embededness of Cobbles)**
5. **Spike Recoveries of Chemicals in Difficult Matrices**
6. **Dissolved Oxygen Concentrations**
7. **5-day Biological Oxygen Demand**
8. **Field Measured Temperatures**
9. **Laser and other New Technologies That Seem to Measure Things Better than Old Technologies (so which answer is most accurate/has the least uncertainty in accuracy?).**

**For such parameters, sometimes precision based estimates of sensitivity are the only ones easily estimated, as explained in more detail below:**

# DQOs for Sensitivity of Field Measurements Using Electronic Probes

**Since one cannot depend on manufacturer's specifications for resolution to be good real-world estimates of field-measurement sensitivity performance, how sensitivity will be estimated in the field, and QC sensitivity objectives should be detailed in the plan. Actual QC performance should be reported as results metadata.**

**The measurement system should be sensitive enough to detect biologically meaningful changes, or any other changes considered significant in trend analyses or various assessments comparing values obtained to values protective of resources of concern.**

**Since the problems with estimating normal detection limits are similar to those encountered in biological or ecological measurements, similar recommendations are given:**

**In common cases where it may be too difficult to estimate measurement sensitivity using signal to noise ratios or standard low level detection limits, SEMI-QUANTITATIVE (presence/absence) measurement sensitivity may be estimated as a limit of detection (LOD). This term is used in general text books, EURACHEM references, and some EPA FIFRA documents. The LOD is the point at which "a measured value becomes...larger than the uncertainty associated with it." The LOD can be defined in a number of ways, such as the background response plus three times the standard deviation of the lowest measurable, concentration, three times the signal-to-noise ratio of baseline noise, and (if all else fails) three times the standard deviation of the lowest measurable concentration " (EPA. 1998. Final report of the FIFRA Scientific Advisory Panel open meeting held in Arlington, Virginia, on March 24-25, 1998, http://www.epa.gov/oscpmont/sap/1998/march/chapb-1.pdf).**

**Quantitative sensitivity may be estimated in one of two ways:**

**First choice, and probably most related to "what we really believe" would be to estimate quantitative sensitivity as a measurement uncertainty interval. Standard NIST/ISO methods to calculate measurement uncertainty are provided in more detail in Section IV-F.2 of this document. The equations include sum of squares, with the squares being squares of standard deviations of precision, systematic error, imperfections in certified reference materials, and any other major contributors to uncertainty. One then takes the square root of the this sum and multiplies it by the middle t value if sample size is less than 30 to come up with a kind of pooled square root to express measurement uncertainty.**

**In situations where neither standard EPA nor USGS detection limits nor NIST/ISO measurement uncertainty are easily estimated (where all else fails, such as for field measurements of pH and oxygen, for example), ten times the standard deviation of repeat measurements (repeatability precision context) of the lowest measurable concentration (or other "signal"), should be used to estimate measurement sensitivity as a limit of quantification (LOQ, see**

**http://www.epa.gov/oscpmont/sap/1998/march/chapb-1.pdf). If it is too hard to determine which samples have the lowest concentrations or signals, just pick some that are relatively low or are typical of the range being measured, and do replicate measures on those.**

**At least eight samples should be re-measured to estimate sensitivity, and the process should be repeated every 20 samples or when something changes (new instrument, new technician, deployment of continuous monitoring probe, retrieval of a continuous monitoring probe, etc.).**

### V-B.4 Detection Limits, A Particular Type of Low-Signal Level Measurement Sensitivity

**Detection limits are the type of sensitivity usually used by chemical laboratories to estimate sensitivity at the lowest concentrations possible with a stated degree of confidence. Detection limits relate to the ability of the measuring system to (confidently) discriminate between measurement responses representing the smallest truly different levels of a variable of interest.**

**Low concentration detection limits are especially important for certain toxic or harmful chemicals that can have effects at even very low levels. In these cases, it is particularly important that the lab can achieve quantitative detection limits that are below the benchmark, criteria, or other threshold levels known to be associated with harmful effects.**

**Although terminology differs, in laboratory chemical analyses, sensitivity as detection limits should be reported as both a quantitative detection limit and a semi-quantitative detection limit. If detection limits are reported in micro-molar (uM) units, they should also be reported in ppm or ppb units more commonly used in the regulatory and environmental monitoring contexts. The wide variety of terms for both semi-quantitative and quantitative detection limits are summarized as follows:**

**Semi-Quantitative detection limits are the lowest levels at which one can determine presence/absence. Common examples of such semi-quantitative limits include:**

1. **EPA's Method Detection Limit (MDL) as defined for the Clean Water Act in 40 CFR Part 136, Appendix B and widely used by States (see Web at http://www.access.gpo.gov/nara/cfr/cfrhtml_00/Title_40/40cfr136_00.html). The MDL is also used for RCRA methods and in SW—846 is defined as "The minimum concentration of a substance that can be measured and reported with 99% confidence that the analyte concentration is greater than zero and is determined from analysis of a sample in a given matrix type containing the analyte" (SW-846 QC document at http://www.epa.gov/epaoswer/hazwaste/test/pdfs/chap1.pdf).**
2. **The USG long-term MDL (LT-MDL, a semi-quantitative or estimated limit, see C. J. Oblinger Childress, W. T. Foreman, B. F. Connor, and Thomas J. Maloney. 1999. New Reporting Procedures Based on Long-Term Method Detection Levels and Some Considerations for Interpretations of Water-**

Quality Data Provided by the U.S. Geological Survey National Water Quality Laboratory, available on the Internet at http://water.usgs.gov/owq/OFR_99-193/.

3. The limit of detection (LOD). This term is used in general text books, EURACHEM references, and some EPA FIFRA documents. The LOD is the point at which "a measured value becomes...larger than the uncertainty associated with it." The LOD can be defined in a number of ways, such as the background response plus three times the standard deviation of the lowest measurable, concentration, three times the signal-to-noise ratio of baseline noise, and three times the standard deviation of the lowest measurable concentration " (EPA. 1998. Final report of the FIFRA Scientific Advisory Panel open meeting held in Arlington, Virginia, on March 24-25, 1998, http://www.epa.gov/oscpmont/sap/1998/march/chapb-1.pdf).
4. The Critical Value, LC and Limit of Detection, LD, were originally defined by Lloyd Currie. Essentially equivalent definitions have been adopted by ISO and IUPAC. LC is the lowest result that can be distinguished from a blank (a semi-quantitative or qualitative (presence/absence) limit.

A quantitative detection limit is typically a two to ten times higher concentration (or other signal value) than a semi-quantitative detection limit such as an MDL. It represents the minimum concentration at which a method or instrument will measure a relatively low value with low uncertainty in accuracy. Typical quantitative detection limits include:

1. A limit of quantification (LOQ), a detection limit that can be defined in a number of ways, such as the background response plus ten times the standard deviation of the lowest measurable concentration, ten times the signal-to-noise ratio of the baseline noise, ten times the standard deviation of the lowest measurable concentration, etc." (EPA. 1998. Final report of the FIFRA Scientific Advisory Panel open meeting held in Arlington, Virginia, on March 24-25, 1998, http://www.epa.gov/oscpmont/sap/1998/march/chapb-1.pdf).
2. Other terms for quantitative detection limits have included minimum levels (MLs), practical quantitation limits (PQL's methods, minimum reporting levels (MRLs) minimum quantitation limits (MQLs), and laboratory reporting levels (LRLs). Whatever the terminology, it should be made clear that what is being discussed or listed is a quantitative detection limit, and how that limit was determined should be listed in the plan.
3. Quantitative values above the "quantitation limit" are classified in EPA's new STORET database as "Detected and Quantified" This is ideal, and according to EPA STORET Staff, this is the only choice which permits reporting a result value. In less than ideal and rare circumstances where the analyte is present above the quantitation limit but is also above the limits of accurate calibration, the value is classified in STORET as "Present Above Quantitation Limit."

4. **In RCRA work, SW-846 QC guidance states that: "The estimated quantitation limit (EQL) is the lowest concentration that can be reliably achieved within specified limits of precision and accuracy during routine laboratory operating conditions. The EQL is generally 5 to 10 times the MDL. However, it may be nominally chosen within these guidelines to simplify data reporting. For many analytes the EQL analyte concentration is selected as the lowest non-zero standard in the calibration curve. Sample EQLs are highly matrix-dependent" (SW-846, Chapter One, Section 5.0, Definitions, at http://www.epa.gov/epaoswer/hazwaste/test/pdfs/chap1.pdf).**
5. **For CERCLA work, MDLs (except for contract required detection limits -CRDLs) are generally similar to those required in 40 CFR Part 136 for the Clean Water Act, but quantitative detection limits are often handled differently. For metals, the MDL has to be less than half the listed contract required quantitation limit (CRQL, http://www.epa.gov/superfund/programs/clp/download/ilm/ilm52d.pdf).**

**Between quantitative detection limits and semi-quantitative detection limits, uncertainty in the measurement is high. The correct inference is typically presence/absence (a non-quantitative estimate). To understand this, let's take the example where the semi-quantitative detection limit (MDL or LOD) is 0.01 ppm and the quantitative detection limit (LOQ) is 0.05. If the true environmental value is 0.01, right at the detection limit, there is a 50% chance that the answer obtained is higher, and a 50% chance that the answer obtained is lower. If one obtained an answer of 0.02, one could not conclude that the concentration at the location was twice as much as the value of 0.04 obtained at another location. That is because between the MDL and LOQ, the values are not quantitatively comparable, and in both cases the proper conclusion is that the analyte seemed to be present but was below the LOQ.**

**Therefore, number value results between the quantitation limit and the semi-quantitative detection limit (MDL) are properly classified in EPA's new STORET database as "Detected, Not Quantified." Values in this range are considered an estimate. Most labs still report a number, and in the Superfund program and some past scenarios therefore a number has been reported, often with a notation that the value is "estimated."**

**Some screening tests (test strips, etc.) do not even produce a number but rather just give an answer of present or absent, In the new STORET database, if the result of the test was "present", this special case is considered "Present Below Quantitation Limit," and no number value are reported.**

**Again, although some labs report number values in this range, typically the confident information value is simply that the analyte is present. Therefore, we would have no confidence that an estimated value of 0.4 would really reflect a larger true environmental concentration than an estimated value of 0.1, if both were in the this range between semi-quantitative and quantitative detection limits.**

**Below semi-quantitative detection limits, we have zero confidence in measurement results (measurement uncertainty is infinite). This would correspond**

**with EPA STORET Code: "Not Detected." No value is recorded, though sometimes one sees "less than" values in reports to provide the semi-quantitative detection limit.**

**Chemical laboratory detection limit concepts and terminology are complex. Those not familiar with detection limit terminology and the differences between EPA's MDL terminology (and alternatives that USGS now considers superior), should probably first read a recent EPA discussion on detection limit issues and terminology [EPA 2001. Guidance on Data Quality Indicators (EPA QA/G-5i) at http://www.epa.gov/quality/].**

**Those trying to compare results with USGS data need a more complete understanding of how USGS calculates detection limits and are encouraged to read USGS's summary discussions on detection limits (C. J. Oblinger Childress, W. T. Foreman, B. F. Connor, and Thomas J. Maloney. 1999. New Reporting Procedures Based on Long-Term Method Detection Levels and Some Considerations for Interpretations of Water-Quality Data Provided by the U.S. Geological Survey National Water Quality Laboratory, available on the internet at http://water.usgs.gov/owq/OFR_99-193/.**

**A key concept to understand about detection limits is that as one attempts to measure lower and lower concentrations (or other signal value) and approaches the quantitative limit, or even goes lower towards the semi-quantitative detection limits, the more uncertainty there is in the concentration (or other signal value) determined.**

**NIST has no definition for chemical lab detection limits but recommends a EURACHEM reference explaining measurement uncertainty at the limit of detection/determination (http://www.measurementuncertainty.org/mu/guide/index.html). This reference explains (appendix f1.1) that at very low concentrations (or other signal value, values near the detection limit), measurement uncertainty increases (sensitivity decreases) due several listed factors:**

1. **the presence of noise or unstable baseline,**
2. **the contribution of interferences to the (gross) signal,**
3. **the influence of any analytical blank used, and**
4. **losses during extraction, isolation or clean-up.**

**The same EURACHEM reference explains that because of such effects, as analyte concentrations drop, the relative uncertainty associated with the result tends to increase, first to a substantial fraction of the result and finally to the point where the (symmetric) uncertainty interval includes zero. This region is typically associated with the practical limit of detection for a given method…Here, the term 'limit of detection' only implies a level at which detection becomes problematic, and is not associated with any specific definition. Among other relevant statements in the EURACHEM reference (op cit., above):**

> **The ISO Guide on Measurement Uncertainty does not give explicit instructions for the estimation of uncertainty when the results are small and**

> **the uncertainties large compared to the results. Indeed, the basic form of the 'law of propagation of uncertainties', described (in NIST/ISO measurement uncertainty equations) may cease to apply accurately in this region; one assumption on which the calculation is based is that the uncertainty is small relative to the value of the measurand. An additional, if philosophical, difficulty follows from the definition of uncertainty given by the ISO Guide: though negative observations are quite possible, and even common in this region, an implied dispersion including values below zero cannot be "... reasonably ascribed to the value of the measurand" when the measurand is a concentration, because concentrations themselves cannot be negative… These difficulties do not preclude the application of the methods outlined in this guide, but some caution is required in interpretation and reporting the results of measurement uncertainty estimation in this region.**

**In a recent dispute, EPA refused to limit the definitions of sensitivity and detection limits to the more narrow sense of signal to noise concepts because signal to noise ratio applications are "limited to specific types of measurement techniques, such as gas chromatography/mass spectrometry (Environmental Protection Agency. 2003. Technical Support Document for the Assessment of Detection and Quantitation Concepts http://www.epa.gov/waterscience/methods/det/dqch1-3.pdf).**

**It should be kept in mind that changes in monitoring methods, including changes in detection limits and censoring methods for values below detection limits, can make trend analyses difficult. Thus, for trend detection, it is preferable that there should be no analytical procedure changes without compelling reason, and networks should be designed with this in mind.**

**However, since some changes in methods are typically inevitable in long term monitoring, networks are told to keep an archive of all the versions of sampling and analyses protocols used. Any method or protocol or analysis changes, and how such changes will change estimates of identical samples, should be in included in archives and databases available those that might attempt to understand trends**

**For continuous monitoring field probes and for many biological and physical habitat variables, the phrase "detection limit" is often not used to document sensitivity.**

## DQOs for Laboratory Detection Limits

**The details of how semi-quantitative and reliably quantitative detection limits are determined should be included in the plan and in metadata reported to STORET. In other words, for laboratory chemical analyses, measurement sensitivity should typically be listed in the plan in following two ways:**

**1) Qualitative or semi-quantitative limits**

**Semi-quantitative limits should be specified for each parameter for which they can be calculated. Whether they are called method detection limits (MDLs) or long term method detection limits (LT-**

MDLs) or something else, if a non-quantitative detection limit is being discussed or listed, how it is calculated should be detailed and it should be made clear whether it is a qualitative detection limit or a semi-quantitative detection limit.

**2) Quantitative detection limits**

**Quantitative limits should be specified for each parameter for which they can be calculated. Whether they are called minimum levels (MLs) or practical quantitation limits (PQL's methods, minimum reporting levels (MRLs) or the limits of quantitation (LOQs or minimum quantitation limits (MQLs), or laboratory reporting levels (LRLs), or something else, how it is calculated should be detailed and it should be made clear that what is being discussed is a quantitative detection limit.**

**As detailed in section V-A.2 (Selecting a Chemical Lab), whenever possible, quantitative detection limits should be below all comparison benchmarks and concern levels such as chronic exposure water quality standards. The plan should provide evidence that the proposed lab measurement sensitivity performance levels will be sufficient to answer specified questions. This need will sometimes drive the lab chosen to do the work, since not all labs will be able to achieve adequately low quantitative detection limits.**

**Ideally the (semi-quantitative) method detection limits should be no more than 20% of the screening value" (Ed Laws, University of Hawaii, Personal Communication, 2003). This would typically insure that the quantitative detection limit was below the screening value, though in some cases in might have to be no more than 10% of the screening value.**

**Unless otherwise justified, the quantitative detection limits shall typically be defined in terms of multiples of the semi-quantitative detection limit or as an LOQ. The quantitative detection limits shall typically be 2 to 10 the semi-quantitative detection limit In laboratory chemical analyses, unless otherwise justified, the quantitative detection shall never be less than 2 times the MDL or LT-MDL.**

**Unless otherwise justified, for regulatory monitoring, the detection limit terminology should be standardized to regulatory agency requirements. Many, but not all of these use EPA MDL limit terminology as defined in 40 CFR Part 136, Appendix B (3 (see Web at http://www.access.gpo.gov/nara/cfr/cfrhtml_00/Title_40/40cfr136_00.html).**

**If one is to compare mostly with USGS data, the type of MDL used could the USGS long-term MDL (LT-MDL, a semi-quantitative or estimated limit). In this case the quantitative detection limit would be a laboratory reporting limit (LRL, calculated as two times the LT MDL, a value at which the false negative probability is limited to 1%) as explained in detail by Oblinger et al. 1999 (op cit., http://water.usgs.gov/owq/OFR_99-193/.**

**In any case, the plan should specify exactly how detection limits will be determined, since method suggested by EPA in 40 CFR Part 136 is different than**

the f-pseudosigma method used by USGS WRD (LT-MDL method) above, and both are different than the pooled standard deviation (pooling variances from a method/reagent blank plus a low concentration sample) method used by the USGS BRD Columbia lab.

The plan should also specify how individual observations will be judged as either useful or not useful, and how they will be treated statistically if judged not useful. Unless otherwise justified, "useful quantitative data" and "useful qualitative data" should be defined consistently with definitions at the end of the appendices.

## DQOs For Censoring Values Below The Detection Limits

How values below detection limits will be handled shall be detailed, and how outliers will be handled shall be detailed in the plan.

General rules of thumb that should be followed include the following (Helsel, D.R. 1990. Less than obvious statistical treatment of data below the detection limit. Environmental Science and Technology 24:1766-1774):

Never delete non-detects. This will produce severely biased results.

Do not substitute 1/2 or some other fraction of the detection limit for non-detects. This may create a signal that was not there in the data, or obscure a signal that was there.

Censor values at either the LT-MDL (detection limit) or the LRL (quantitation limit) according to your judgment of best practice, but don't use both.

Do not report values between the limits (and so be censoring at the LT-MDL), but report nondetects as <LRL, for example. This is called "informative censoring" in the literature, and introduces bias.

In many cases using nonparametric tests get around the censoring problem. So to compare two groups (with one detection limit) a Mann-Whitney test works perfectly well. A t-test may run into big trouble depending on what is substituted for nondetects. The nonparametric tests don't require anything to be substituted.

The best methods for handling nondetects are those that use the values for the detected observations plus the probabilities of being at or below each detection limit for censored values.

However, some of these probability-based more robust methods are complex and are not as uniformly needed in simple precuationary comparisons with standards and threshold levels as they are in complex trend assessments. Using the

**precautionary principle in comparisons with standards or benchmarks, one can typically justify censoring values at quantitative detection limits:**

## COMPARISONS WITH STANDARDS OR OTHER BENCHMARKS:

**In order to be protective of important NPS resources, for comparison with water quality standards or other protection benchmarks, all data below quantitative detection limits [typically called minimum levels (MLs), or practical quantitation limits (PQL's), or minimum reporting levels (MRLs), or the limits of quantitation (LOQs), or minimum quantitation limits (MQLs), or laboratory reporting levels (LRLs) should be censored to the quantitative detection limits before calculations of averages or confidence intervals.**

**This may seem surprising to some but is consistent with USGS guidance for certain scenarios (C. J. Oblinger, Childress, W. T. Foreman, B. F. Connor, and Thomas J. Maloney. 1999. New Reporting Procedures Based on Long-Term Method Detection Levels and Some Considerations for Interpretations of Water-Quality Data Provided by the U.S. Geological Survey National Water Quality Laboratory, available on the internet at http://water.usgs.gov/owq/OFR_99-193/data.html#effect):**

> **"When examining an individual concentration as a basis to assess compliance with environmental regulations, a high degree of certainty is needed to account for the possibility of reporting a false positive or false negative concentration. That is, the user does not wish to report that a potential contaminant is present when it is not or that the analyte is absent when, in fact, it is present. The user may choose to increase data certainty by ignoring estimated concentrations and censoring all data at the LRL, at a higher historic MRL, or at any other project-specific level that is greater than the LRL."**

**Censoring to quantitative detection limits, though not often optimal or even desirable for trend assessment, is appropriate when a data user is assessing compliance with environmental regulations or trying to determine if a concentration or summary statistic is below a concern threshold for endangered species or other special-value resources in National Parks. If such especially high value resources are at risk, it is appropriate to apply the precautionary principle and avoid false negatives (saying the contaminant is below a certain concentration, when it really is not).**

**As pointed out by USGS, a user that first censored data to the quantitative LRL detection limit "may choose to re-censor these data to the LT–MDL (a lower concentraton value, semi-quantitative detection limit), thereby accepting the increased risk of false-negative error (up to 50 percent, see C. J. Oblinger, Childress, W. T. Foreman, B. F. Connor, and Thomas J. Maloney. 1999. New Reporting Procedures Based on Long-Term Method Detection Levels and Some Considerations for Interpretations of Water-Quality Data Provided by the U.S.**

**Geological Survey National Water Quality Laboratory, available on the internet at http://water.usgs.gov/owq/OFR_99-193/data.html#effect).**

**Censoring to half the MDL would result in a still higher risk of false negatives than censoring the data to the MDL.**

**Likewise, the user may choose to censor historical data in data reports at the new LRL (a quantitative detection limit) when a high degree of certainty is required (http://water.usgs.gov/owq/OFR_99-193/summary.html#summary).**

**EPA has recognized the validity of censoring to detection limits in some situations: EPA agrees that the European Union approach of censoring to detection limits, which yields a "worst-case" (precautionary principle, highest possible) estimate of the pollutant concentration, can serve as a useful regulatory tool for encouraging the analytical and regulated community to pursue measurements at the lowest levels necessary to protect human and ecological health. However, EPA also cautions that this approach also should be recognized as a regulatory strategy that effectively censors measurements made below the MDL (Environmental Protection Agency. 2003. Technical Support Document for the Assessment of Detection and Quantitation Concepts, Section 3.1.4 Recovery Correction, page 37, http://www.epa.gov/waterscience/methods/det/dqch1-3.pdf).**

> **Note: Replacing nondetects with the quantitative detection limit will give the highest mean value, but not necessarily the highest or lowest standard deviation, and so test statistics built on these (t-tests) won't necessarily be either conservative or liberal. These issues can be complex which is why I am explaining them more in a book planned for publication in 2004. In the meantime, there are some guidelines for interpreting censored environmental data at www.practicalstats.com (Dennis Helsel, USGS, Personal Communication, 2003).**

**Observations between the semi-quantitative MDL and quantitative detection limits should be judged qualitatively (semi-quantitatively) as present. However, only values above the quantitative detection limit (or censored to that value) should be judged as ideally appropriate for precautionary principal comparisons with water quality standards and other benchmarks. In cases where the precautionary principle for comparison with standards or benchmarks is not the overriding value, more sophisticated methods for handling non-detects, such as those discussed below for trend assessment, should be considered.**

## TREND ASSESSMENT

**Replacing data below detection limits with detection limits or half the detection limit should not be done when assessing trends in data. Instead more robust, sophisticated, and trend-appropriate methods of dealing with non-detects should be utilized.**

**As long term monitoring proceeds over time, lower detection limits are often used. Then if values below the detection limit are censored to the limit or half the limit, an apparent trend can appear when really the only change was in detection**

**limits. Data censoring is of concern when using non-parametric statistics (such as the popular Seasonal Kendall and Mann-Kendall tests for trends) because of the production of multiple tied values (D.G. Smith and P.B. McCann. 2000. Water quality trend detection in the presence of changes in analytical laboratory protocols. Paper Presented at the National Water-Quality Monitoring Council (NWQMC) National Conference 2000, Accessed on Internet November 2003 at http://www.nwqmc.org/2000proceeding/papers/pap_smith(b).pdf). This paper gives strong arguments and examples of why:**

1. **One should not change detection limits as the long term monitoring proceeds, and**
2. **One should not use simple substitution methods, such as censoring to a detection limit or half a detection limit) in trend assessment, and**
3. **Values between the quantitative and semi-quantitative detection limits, though higher in uncertainty than those above the quantitative detection limit, should be used in trend assessment. This advice is opposite that given for precautionary principle comparisons with standards or other benchmarks or thresholds for harmful effects.**

**More sophisticated and trend-applicable methods for handling non-detects often consider the distribution of data above the censoring level. That distribution is then often used to extrapolate values below the censoring level, providing the advantage that the distribution is based on observed data rather than assumed. However, both the distributional methods and robust methods work poorly with small sample sizes. For more details, see**

1. **USGS guidance on the Internet (USGS Open File 99-193 at http://water.usgs.gov/owq/OFR_99-193/data.html), and**
2. **Helsel, D.R. and R.M. Hirsch 1992. Statistical Methods in Water Resources. Studies in Environmental Science 49, Elsevier Publishing, NY, http://water.usgs.gov/pubs/twri/twri4a3/pdf/twri4a3.pdf.**

### V-B.5 Measurement Resolution

**Resolution is a term used by makers of field measurement meters but not typically seen in environmental quality assurance project plans (QAPPs). For most applications, use of the word resolution is not justified, since those using the word resolution are often talking about other more commonly understood QC concepts, such as precision, uncertainty in accuracy, sensitivity, detection limits (a special case of sensitivity when signals are low), or measurement uncertainty. Since these concepts are defined in detail separately herein and more universally understood, there is typically no need to address "resolution" separately in detailed study plans or QAPPs.**

**Given the variability and hostility of field conditions, assuming that lab-derived "resolution specifications" are good real-world estimates of detection limits [the most common way sensitivity is controlled and a special case (low-level signal)**

**of measurement sensitivity] or other more general forms of measurement sensitivity, would be wrong in most cases.**

**When the word resolution is used, it always seems to relate to the fineness of the measurement scale. However, that is about the only safe generalization.**

**To the surprise of some, the word resolution is often (counter-intuitively) not necessarily linked to an ability to accurately "resolve" small differences with a stated amount confidence. Unlike the word sensitivity, the word resolution is often used in the context of no implied capability of the measurement system to accurately "resolve" between signals vs. noise. Often, but not always, it is just the number of decimal places/significant figures displayed by the meter.**

**In this context, if one electronic meter displayed 5 digits for pH (say a pH of 7.2598) that meter would have a different resolution than a competitor's meter that displayed a pH of 7.3, but in neither case would the resolution have anything to do with uncertainty in accuracy. Uncertainty in accuracy relates to sensitivity, not to resolution. Unlike sensitivity, there is not necessarily a defined degree of confidence or uncertainty associated with resolution.**

**In a few cases, meter makers link resolution to precision without regard to real-world uncertainty in accuracy or measurement sensitivity. In those cases, they should probably be using the word precision instead of the word resolution.**

**NIST has no definition for resolution. However, a publication recommended by NIST, The International Vocabulary of Basic and General Terms in Metrology (called the VIM, published by ISO 1993) defines (VIM 5.12) "resolution (of a displaying device)" as the smallest difference between indications of a displaying device that can be meaningfully distinguished." The same definition is provided by EURACHEM (http://www.measurementuncertainty.org/mu/glossary/index.html), a reference also recommended by NIST (Tyler Estler and Steve Phillips, NIST, Personal Communication, 2003). Although the word "meaningfully" seems to hint at some degree of confidence, resolution is used in varied and often conflicting ways. To avoid confusion, it is suggested that the word resolution not be used unless clearly defined and clearly differentiated from other concepts typically controlled in QAPPs, including sensitivity, measurement uncertainty, detection limits, and precision.**

**Resolution is used in relationship to confidence or precision less often than the word sensitivity or the phrases detection limits or measurement uncertainty. It is often unclear what "resolution specifications" really mean, so check with the manufacturer of the meter and/or don't use the term.**

**In the past, rounding rules for final results (only the last digit reported are supposed to have any uncertainty), and perhaps rounding results based on meter specifications for resolution, have been ways utilized to at least very crudely factor in uncertainty in a reported final results. However, measurement uncertainty is a better way to consider uncertainty.**

**An example of a problem with trying to use "resolution specifications" in the context of continuous, long-term deployment of multi-depth probes in lakes concluded there were real-world differences in what "they believed" vs. "resolution specifications" provided by the manufacturers (Remote Underwater Sampling System/RUSS Quality Assurance Summary, University of Minnesota Duluth,**

**Natural Resources Research Institute Water Educational Website at http://wow.nrri.umn.edu/wow/under/qaqc.html, see also http://lakeaccess.org/QAQC.html):**

Table 2. Reporting limits for RUSS sensor data (Hydrolab or YSI sensors)

| Depth (m) | Temp (oC) | DO (mg/L) | DO % saturation | pH | EC (uS/cm) | Turbidity (NTUs) |
|---|---|---|---|---|---|---|
| Resolution (what is reported by the RUSS sensors) | | | | | | |
| ± 0.12 | ± 0.1 | ± 0.1 | ± 0.1 | ± 0.1 | ± 1 | ± 1 |
| Estimated Accuracy (what we really trust) | | | | | | |
| ± 0.3 | ± 0.15 | ± 0.2 | ± 2 | ± 0.2 | ± 10 | ± ~3 |

**In this context, RUSS specialists were considering resolution to be "the smallest reading shown for a particular parameter is likely to be considerably lower than the error associated with differences in time, with depth fluctuations, and with sensor drift and calibration accuracy.**

**This table uses the manufacturers "resolution specifications" rather than their accuracy specifications, but none of the specifications provided by manufacturers necessarily relate well to real-world uncertainty in accuracy or to measurement sensitivity.**

**Do we really care what the ideal (controlled) lab condition specifications for resolution or accuracy are? Don't we care more about real world uncertainty in accuracy (measurement uncertainty). Isn't that a better estimate of what we really believe?**

**In today's world, more modern, defendable, and consistent ways to account for uncertainty include:**

1. **Calculating measurement uncertainty per NIST/ISO guidance (see Section IV-F.2) for values in the range where quantitative measurements are possible.**
2. **Using quantitative detection limits and semi-quantitative detection limits to explain the amount of uncertainty associated with low signal strength values, values below quantitative detection limits (section V-B.4).**
3. **In cases where the above two are not practical or applicable, use other EPA-recommended methods related to precision to estimate measurement sensitivity, such as estimating the limit of quantification (LOQ) in the easiest of the following optional ways:**

> **"The LOQ can be defined in a number of ways, such as the background response plus ten times the standard deviation of the lowest measurable concentration, ten times the signal-to-noise ratio of the baseline noise, ten times the standard deviation of the lowest measurable concentration, etc." (EPA. 1998. Final report of the FIFRA Scientific Advisory Panel open meeting held in Arlington, Virginia, on March 24-25, 1998, http://www.epa.gov/oscpmont/sap/1998/march/chapb-1.pdf).**

**The International Union for Pure and Applied Chemistry points out that if sensitivity is to be a unique performance characteristic, it must depend only on the measurement process, not upon scale factors (http://www.iupac.org/goldbook/S05606.pdf). About the only safe generalization one can make about resolution is that it always seems to relate to scale factors.**

**In most cases, resolution and sensitivity are best understood as two different concepts. In the narrow sense, sensitivity relates to the ability to pick out a signal from background noise, whereas resolution is sometimes used simply to imply the fineness of the measurement scale (the number of digits or decimal points displayed), whether or not signals can be correctly differentiated (resolved) from noise at any given fineness of scale (resolution).**

**Certain GIS/Remote sensing and non-linear biological categorization applications may be an exception.**

**In remote sensing, the word resolution is often used for a concept more broadly recognized in other disciplines as sensitivity. For example, in a geospatial guidance document, EPA talks about mapping and remote sensing "resolution" as pixel size in one place and as a "detection limit" (sic) in another. The same document refers to map resolution as "the accuracy with which the location and shape of map features are depicted for a given map scale" in a third areas. Projects collecting considerable new geospatial data (derived from remote-sensing, mapping, and surveying technologies) now have guidance for how QA/QC for geospatial work should been implemented. In such a guidance document, EPA put "detection limit" in parentheses after the word resolution. (EPA 2003, Guidance for Geospatial Data Quality Assurance Project Plans (QA/G-5G), EPA/240/R-03/003 http://www.epa.gov/quality/qs-docs/g5g-final.pdf).**

**Thus, in the geospatial/GIS world, the word resolution is used for concepts that others would recognize as sensitivity, signal to noise ratios, or detection limits.**

**One also sees the word resolution used related to non-linear taxonomic classifications and other non-linear biological or ecological separations into groups.**

**When the emphasis is on correctly categorizing values into groups, the correct or expected answer is not always easy to document, nor is it easy to document signal to noise ratios or other estimates of sensitivity similar to those for physical or chemical measures. In some such scenarios, the concept being communicated is simply fineness of scale rather than sensitivity to a signal. Some recommend using the word "resolution" for these types of situations,**

**For example, in biological work, Wayne Landis uses the word resolution related to non-linear determinations such as levels of taxonomic resolution, levels typically picked to achieve "taxonomic sufficiency" in information content. Wayne states that "One of the reasons that dealing with taxonomy is not identical to working with a chemical or physical measurement is that the classifications that we use in taxonomy are fairly arbitrary. For instance, order, family, genus, and even species are fairly arbitrary notions of hierarchy. Biologists tend to be splitters when the taxonomy deals with organisms that they deeply care about. There are no specific molecular or evolutionary criteria for the designations of genus or family (or even order) so that a direct comparison between different types of organisms is really not possible. In this situation, we do have a linear measurement scale (Wayne**

**Landis, Institute of Environmental Toxicology, Huxley College of the Environment, Western Washington University, Personal Communication, 2003). For a more detailed discussion of the Landis ideas on resolution, see appendix V-B.5.**

**Others have also used "taxonomic resolution" in a similar "fineness of scale" context, for example, Will Clements of CSU stated that the ultimate decision concerning taxonomic resolution is most often determined by logistics and costs. For certain groups (especially chironomids and oligochaetes) species- or even genus-level identification is problematic. Assuming limited resources, the question really becomes should you spend these resources on identification of rare midges to species or should you relax taxonomic resolution and spend your money sampling more sites or sampling more frequently. Sampling more sites or more frequently can add information on variability that may be more valuable than spending the additional resources on a lower level of taxonomic resolution (identifying everything to species level). Some work in marine systems suggest that phyla-level identification can reveal signal to noise (sensitivity) differences, while others working in freshwater systems have argued that species level identification is essential in biomonitoring studies. Our own work suggests that at least for western streams, the appropriate level of taxonomic resolution depends on which group you are assessing and, more importantly, the spatial scale of the investigation (Will Clements, Personal Communication, 2003, for more information see: Clements, W. and M. C. Newman, 2002. Community Ecotoxicology, John Wiley & Sons, 350 pp., http://www.wiley-vch.de/publish/dt/books/bySubjectNU00/bySubSubjectNU/0-471-49519-0/?sID=d05b).**

**For more information on taxonomic sufficiency, see discussions of screening vs. definitive methods in section IV-F.1.**

## DQOs for Measurement Resolution:

**The word resolution should not be used in detailed study plans or QAPPs except for specialized applications such as:**

1. **Geospatial/GIS mapping, or**
2. **Non-linear categorization into groups such as taxonomy and other biological/ecological categorizations.**
3. **Fineness of the measurement scale with no implied degree of measurement sensitivity or uncertainty in accuracy.**

**When it is used, the word resolution shall be defined, given numerical limits to the degree possible, and clearly differentiated from other QC terms such as those discussed below.**

**For other applications, use terms more widely understood and detailed elsewhere herein should be used rather than the term resolution. For example:**

**When the concept being discussed is really the data quality indicator of measurement sensitivity, rather than resolution, measurement sensitivity should be documented as explained in the sections (section V-B.2-4 above).**

**When the concept being discussed is really (uncertainty in) accuracy after factoring both precision and systematic error, rather than resolution, measurement uncertainty should be documented as explained in Section IV-F.2.**

**If neither detection limits nor measurement uncertainty are easily obtained or estimated, other (more general) forms of measurement sensitivity, such as an LOQ, rather than "measurement resolution," shall be specified as explained in Section V-B.2 to 3.**

**When the concept being discussed is really measurement precision, rather than resolution, measurement precision should be documented as explained in the sections V-D and V-E.**

### V-B.6 Calibration

**Optimal instrument calibration is typically necessary to achieve optimal measurement sensitivity and reasonably low uncertainty in overall measurement accuracy when one is using an electronic meter or other complex instrument in the measurement process.**

**Calibration of complex measuring instruments is typically done according to manufacturer's specifications unless a more rigorous calibration is needed for regulatory or data comparability reasons. For example, the USGS field manual specifies some calibration SOPs for parameters like field measured pH.**

**Again, there is a strong relationship between the concepts of measurement sensitivity, calibration, and the concept of uncertainty. This is especially true when the topic being discussed is uncertainty related to calibration with calibration grade standards, or uncertainty in a calibration setting. Among the important concepts summarized by NIST (S. D. Phillips, W. T. Estler, T. Doiron, K. R. Eberhardt, and M. S. Levenson. 2001. A Careful Consideration of the Calibration Concept. J. Res. Natl. Inst. Stand. Technol. 106(2), 371–379, http://nvl.nist.gov/pub/nistpubs/jres/106/2/j62phi.pdf) are the following:**

1) **The relationship between measured or indicated values and those of the reference values is a key issue with regards to calibration.**
2) **A traceable measurement requires both an unbroken chain of comparisons back to a reference value and also an uncertainty statement.**
3) **The calibration results are valid for a specified set of validity conditions which may (or may not) be the same as the conditions in the measurand definition.**

4) **An uncertainty statement is associated with a measurement result, not with the measurement instrument (although the instrument is an uncertainty contributor).**

**Obtaining optimal or needed sensitivity typically depends highly on calibration and maintenance. Recommendations for marine/estuarine environment maintenance and calibration of continuous monitoring probes and data loggers (including conductivity, temperature, and depth/CTD water column data loggers and various pH, oxygen, and SC/salinity probes) are found in EPA. 2001. Environmental Monitoring and Assessment Program (EMAP): National Coastal Assessment Quality Assurance Project Plan 2001-2004. United States Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Gulf Ecology Division, Gulf Breeze, FL. EPA/620/R-01/002. (http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf).**

**For details on field measurement calibration of probe-measured field measures (such as (pH, conductivity, dissolved oxygen, and temperature) see separate guidance in Part C at http://science.nature.nps.gov/im/monitor/protocols/wqPartC.doc.**

## DQOs for Calibration of Electronic Instruments

**The plan should document that the measurement-instrument calibration to be used is rigorous enough to ensure measurements accurate enough (uncertainty in accuracy is low enough) for the intended uses and comparable enough to make the data comparable to regional regulatory or status and monitoring data sets anticipated to be used for comparison.**

**Measurement calibration methods should be detailed in the plan and in STORET as metadata and should in all cases be good enough to allow measurements to meet the data quality objectives and QC performance criteria listed in the plan.**

**Acceptance criteria for calibration check standards should be specified by the method or SOP. The criteria depend on the parameter. For example, for metals in water, acceptable recovery is typically 90-110%. An acceptable general default, when criteria are not specified in the method or SOP, is ± 15%. When acceptance criteria are not met, the calibration check must be repeated. If the criteria are again not met, the primary analyst must find the source of the problem before proceeding with analyses (http://www.epa.gov/region04/sesd/asbsop/asb-loqam.pdf).**

**Measurement instrument calibrations should be done within the range of the expected field concentration, to ensure the applicability to the measurement target(s) within the applicable matrix (http://wi.water.usgs.GOV/pmethods/PBMS/position/dog.version5.2.html).**

**Why is this important? It is important because different qualitative and quantitative detection limits may apply, and the results may differ to some matrices and not others, due to interferences.**

**Thus, for PBMS and MQO processes to be effective, when buying or preparing a "known value" spike, the spike should be prepared in a matrix similar to the environmental sample rather than just by diluting a known amount of contaminant in pure reagent water (J. Diamond et al. 2001. Towards a definition of performance-based laboratory methods. A position paper of the National Water Quality Monitoring Council Methods and Comparability Board, Technical Report 01-02, Web: http://water.usgs.gov/wicp/acwi/monitoring/nwqmc).**

**For field measurements, final calibration and especially that final check for systematic error should be done in the same environment where the measurement is made and not too long before measurement. In other words, to the degree possible, environmental conditions (temperature, wind, etc.) should be the same when measuring a known reference value for calibration or systematic error (bias) purposes, as when measuring environmental samples.**

**To achieve data comparability with USGS data, one should optimally calibrate the way USGS specifies. Likewise, if one is trying to achieve data comparability with EPA EMAP-Marine protocols, one has to follow their recommendations for the minimum number of standards used in instrument calibration  As pointed out above, for field measurements of pH, the USGS recommends 2-3 points be used in calibration, DOE 2, EMAP estuarine program recommends 2,and some lab experts recommend 6 for certain critical applications such as global warming-related measurements of seawater. In any case, after one has finished final calibration steps, one should then check instrument performance against an additional (new) NIST traceable or other very high quality certified reference material standard to get an estimate of systematic error.**

**Chemical lab calibration should meet the following criteria:**

> **Regulatory monitoring calibration should be equal to or better than calibration requirements of the regulatory agency.  For example, calibration of metals lab measurement for Superfund sites should follow calibration protocols listed by EPA (http://www.epa.gov/r10earth/offices/oea/fginorg.pdf).**
>
> **General trends monitoring should ordinarily use standard USGS calibration methods or superior methods.**

**Like other method details, changes in calibration methods should be kept to a minimum to help ensure long-term data comparability.**

**Studies in which FISH TISSUES are analyzed for contaminant concentrations should use detection limit and calibration performance standards at least as rigorous as that suggested by EPA (EPA, 2000, Guidance for assessing chemical contaminant data for use in fish advisories, Volume 1: Fish Sampling and Analysis - Third Edition. EPA 823-B-00-007 at http://www.epa.gov/ost/fishadvice/volume1/v1ch8.pdf).**

### V-B.7 Statistical/Study Design Sensitivity

**The discussions above in sections V-B.1-6 have all related to the measurement level (measuring a single data point) issues. Sensitivity at a higher level of organization or concern also needs to be considered. This higher level involves the sensitivity of the overall statistical study design to detect important changes in summary statistics (such as mean values). Whereas measurement sensitivity relates to even a single data point, study design or statistical sensitivity relates to the ability to detect an important change in the statistics that summarize multiple data points.**

**Once uncertainty in sensitivity is taken into account, one should determine and control the sensitivity of the statistical study design to limit both false negatives and false positives to an acceptably low level. This is typically done by making sure statistical power and general study designs are both adequate to detect changes considered biologically or ecologically relevant before the study or monitoring begins (for more detail, see section IV-C on inequivalence testing and hypothesis testing).**

**In considerations of statistical/study design uncertainty, if (measurement uncertainty-adjusted) confidence intervals do not overlap and one has adequate sample sizes to assure statistical power of say 95 or 99%, one generally assumes the samples are statistically different**

**Detection limits for biological metrics should typically be based on (beyond bio-equivalent) sensitivity and power analyses (Barbour, M.T., J. Gerritsen, B.D. Snyder, and J.B. Stribling. 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, Second Edition. EPA 841-B-99-002. U.S. Environmental Protection Agency; Office of Water; Washington, D.C., (http://www.epa.gov/owow/monitoring/rbp/ch04main.html#Section 4.4).**

## DQOs for Statistical/Study Design Sensitivity in Field Estimates of Biological or Physical Habitat Condition

**For biological and physical habitat estimates, the minimum change detectable, given the statistical and study design and estimated variability (an also taking into account measurement sensitivity and uncertainty, should be identified in the plan. The study should be designed in a manner that allows detection of effect sizes or changes (trends) considered biologically or ecologically significant**

**Measurement uncertainty should be considered as part of overall uncertainty. If confidence intervals are being compared, typically they need to be increased to account for measurement uncertainty (see section IV-F.2.**

## V-C Data Completeness

**Completeness is a QC data quality indicator usually required to be considered in state and federal aquatic quality assurance project plans to ensure data credibility.**

**Completeness can be expressed as a percentage.**

**Minimum sample sizes needed to answer questions will be driven mostly by other factors required by the Outline for Vital Signs Monitoring Plans (Outline for Vital Signs Monitoring Plans, 2003, http://science.nature.nps.gov/im/monitor/docs/monplan.doc) and mentioned above in earlier steps:**

1. **The need to be able to detect certain threshold values or trigger points as discussed above (section III-V, above).**
2. **An understanding of the variability in the various strata and how the sampling scheme will insure that the value obtained will be representative of the target population being studied (Section IV-A, above).**
3. **The length of sampling and when sampling is no longer representative (Section IV-A, above) of the original question or strata due to changed conditions.**

**Once the minimum sample sizes are established for each vital sign, they should be listed in the QC SOP included in each protocol, typically as part of a table that also summarizes data completeness goals for each vital sign. These data completeness goals are typically given as percentages and developed by working backwards from best professional judgment estimates of the number of projected samples that are not likely to produce usable data, recognizing that it is very rare that 100% of all planned samples are successfully obtained and also pass all data acceptance criteria (such as the QC measurement quality objective QC goals discussed in sections below).**

**If the statistical power analysis indicates that a minimum of 100 samples are needed to determine whether or not a threshold or trigger point effect size has been achieved, and if one estimates that perhaps 20% of the planned samples will not produce useable data, one typically needs to plan for more samples (say 120 or 130) in recognition that some samples will not produce usable data due to real world factors such as meter failure, loss of equipment in a flood, data that does not meet QC measurement quality objective acceptance criteria, etc. Once this has been done, a reasonable data completeness QC indicator goal (such as 80%, 90% or whatever value is appropriate for the issue at hand) should be provided for each vital sign as part of a QC SOP (for more information on data completeness, see appendix V-C).**

**Although this detailed information will be provided in the QC SOP in each protocol, the monitoring plan itself should indicate that data completeness was considered for each protocol, and should point readers to the applicable protocol QC SOPs for more details.**

### Generic NPS Completeness Data Quality Objective:

**Since this is long term monitoring, a failure of the samples in one sampling trip or even one year may not be a fatal flaw, so no pre-project QC performance standard (like 70% of the data must be complete) is required before monitoring begins. However, a percentage of the data considered complete should be reported to the monitoring network at least once per year, and if more than 20% of the data is not considered useful (see definitions section), the monitoring network should take steps to remedy problems.**

**Typical Data QUANTITY Objective Related To Completeness.**

**The plan should document that planned data quantities will be sufficient to answer questions; even though 100% completeness is not achieved after sampling is finished.**

**Note: What is the minimum amount of sufficient quality data needed to make the decisions, considering that data planned seldom turns out to be 100% "complete?" For more details on the data quality indicator "completeness", see appendix V-C as well as the latest proposed EPA guidance in Section 4.2 of EPA 2001. Guidance on Data Quality Indicators (EPA QA/G-5i) at http://on-linelearning.ca/idec4433/epaqaqc2000/g5i-prd.pdf).**

## Typical Data Quality Objectives for Completeness of QUANTITATIVE Chemical Data:

**Unless otherwise justified, useful quantitative data should not include values that do not meet the definition for useful quantitative data at the end of this document.**

**Completeness publication criteria for continuous monitoring of temperature, pH, specific conductance, dissolved oxygen, and turbidity should follow the suggestions of USGS (Wagner, R. J., H. C. Mattraw, G. F. Ritz, and B. R. Smith, 2000. Guidelines and Standard Procedures for Continuous Water-Quality Monitors: Site Selection, Field Operation, Calibration, Record Computation, and Reporting. U. S. Geological Survey, Water Resources Investigations Report 00-4252, http://pubs.water.usgs.gov/wri004252) unless otherwise justified.**

**Once per year during long-term monitoring, a calculation of quantitative completeness should indicate that degree of data completeness (measured for each set of data by dividing the number of useful quantitative measurements actually obtained by the number of measurements that were planned) should be sufficient to generate enough useful data to answer quantitative questions with sufficient statistical power. What percentage of non-useful data would result in an inadequate number of samples in individual strata or other defined sample units? If periodic**

**checks of completeness indicate that an insufficient number of useful observations are being made to answer the questions identified, the study design should be modified to correct the deficiency.**

## Typical NPS Completeness Data Quality Objectives for QUALITATIVE or SEMI-QUANTITATIVE Chemical Data:

**Once per year during long-term monitoring, a calculation of qualitative completeness should indicate that degree of data completeness (measured for each set of data by dividing the number of useful qualitative (or semi-quantitative) measurements actually obtained by the number of measurements that were planned) should be sufficient to generate enough useful data to answer qualitative questions.**

**Useful qualitative data should not include values that do not meet the definition of useful qualitative data in the definitions section at the end of the appendices.**

**Qualitative or semi-quantitative data should not be used for trend analyses or to try to demonstrate that recorded values are above or below water quality standards or other benchmarks, since the chance for false negatives is too great (see discussion of detection limits in section V-B.4).**

## QC Measurement Quality Objectives

**As of 2001, federal agencies such as EPA, USGS, NOAA, and methods standardization groups such as the Methods and Data Comparability Board (MDCB) Accreditation Workgroup of the federal (interagency) National Water Quality Monitoring Council have all endorsed the need for validated methods and concise, achievable performance criteria for measurement quality objectives and data quality indicators (J. Diamond et al. 2001. Towards a definition of performance-based laboratory methods. A position paper of the National Water Quality Monitoring Council Methods and Comparability Board, Technical Report 01-02, Web: http://water.usgs.gov/wicp/acwi/monitoring/nwqmc. The NPS endorses these same concepts. Measurement quality objectives (MQOs, also sometimes referred to as "quantitative QC performance criteria" or Quality control measurement performance standards) are recommended for the following measurement quality indicators such as precision and, systematic error (bias).**

### Introduction to Precision:

**As made clear by both NIST and ISO, the first thing to determine when one is considering the concept of precision is to identify the type of precision one is discussing, precision in the context of repeatability, or precision in the context of**

**reproducibility. Both NIST (Document 1297) and ISO [ISO 3534-(D.2)] consider precision to mean "the closeness of agreement between independent test results obtained under stipulated conditions" (N. Taylor and C. E. Kuyatt. 1994. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results NIST Publication TN 1297 (http://physics.nist.gov/Document/tn1297.pdf). The key question is: Are those stipulated conditions repeatability, where nothing in the measurement process changes, or reproducibility, where something in the measurement process has changed? If the sample being measured has changed rather than something in the measurement process changing, the concept being discussed may be true variability (reflecting sample heterogeneity) rather than (random error) lack of precision as either reproducibility or repeatability reflecting noise in the measurement process.**

**Both ISO and NIST view the concept of precision as encompassing both repeatability and reproducibility, since ISO defines repeatability as "precision under repeatability conditions," and reproducibility as "precision under reproducibility conditions." Nevertheless, precision is often taken to mean simply repeatability." To help avoid confusion that has been rampant, NIST makes a "strong recommendation that such terms not be used as synonyms or labels for quantitative estimates. For example, the statement 'the precision of the measurement results, expressed as the standard deviation obtained under repeatability conditions, is 2 standard deviation units' is acceptable, but the statement "the precision of the measurement results is 2 standard deviation units" is not."http://physics.nist.gov/Document/tn1297.pdf). See also NIST compatible ISO terminology in ISO Glossary (http://www.westgard.com/isoglossary.htm).**

**Measurement precision under repeatability conditions is the variability of repeated measurements of the same thing, when no factor, other than perhaps a very short period of time between measurements, is changed. It is the closeness of agreement (expressed as variability) of repeat measurements (between independent test results obtained under prescribed stipulated conditions). In practice in water quality, contaminants or biological studies, it is a measure of the variability in results when one is measuring the same (homogenous) thing repeatedly.**

**Standard Methods defines repeatability as intrinsic measurement variability as reported in standard deviation units after measuring the same thing multiple times within one lab, using only one operator, using only one instrument, on only one day. This is often identical to (or close to identical) to how many would measure precision, so in many cases when just sees the word precision, the concept is precision as repeatability. Thus many tend to see precision and precision as repeatability as synonyms or very close to it.**

**Precision in the context of reproducibility, on the other hand, is a totally different concept, typically pertaining to repeated measurements where one or more measurement process factors (day, operator, lab, etc.) are varied (American Public Health Association, American Water Works Association, and Water Environment**

**Federation. 1998. Standard methods for the examination of water and wastewater, 20th Ed. American Public Health Association, Washington, D.C., see also consistent definitions of NIST at http://physics.nist.gov/Document/tn1297.pdf).**

**If one quickly measures exactly the same thing, say the width of a desk, twice, that would be a "duplicate" trial for measurement precision under repeatability conditions.**

**On the other hand, reproducibility precision is a sound science and data comparability basic, and it also rightly tends to be stressed in information quality guidelines. If a lab doing work for the NPS analyzes a water sample and gets an answer of 0.05 mg/L total mercury, another lab approved by the same federal agency (EPA/NELAP, NOAA, FWS, USGS, etc.) after round robin checks (see lab selection discussion in step V-A.2) that has similar data acceptance performance standards for repeatability precision (duplicates) and for systematic error (bias) control should also get an answer very close to 0.05 mg/L for that same sample. The main way this type of QC is covered herein is in the data comparability (step V-A) and lab selection (step V-A.2) steps. Reproducibility precision is less straightforward for biological and habitat measures, so for those types of parameters, additional efforts should probably be directed at bounding uncertainty in data quality and comparability related to reproducibility precision.**

**Where good alternatives are not available, some have argued that measuring two different samples, perhaps those from the same general area of soil, or two river samples from the same spot rapidly (little time between samples) is one form of "field measurement precision." However, logic would dictate that such measurements might really be estimating true variability (reflecting true heterogeneity in the samples) in two different samples rather than precision under repeatability conditions, since the samples being measured may not be exactly the same. Thus care must be taken in the terminology used and in the interpretation of the results. If one really wants to estimate measurement precision under repeatability conditions, one may have to thoroughly mix and homogenize the two river or soil samples to be sure the correct answer should be the same, before a duplicate measure of precision under repeatability conditions is made.**

**In a document on QC design in NAWQA, USGS uses the word variability rather than the word precision for the concept of precision under reproducibility conditions, clarifying that "variability is the degree of random error in repeated measurements of the same quantity" (http://water.usgs.gov/nawqa/protocols/OFR97-223/ofr97-223.pdf). USGS further clarifies that "Variability is the degree of random error in independent measurements of the same quantity and is the opposite of precision--the degree of mutual agreement" (J.D. Martin, 2002. Variability of Pesticide Detections and Concentrations in Field Replicate Water Samples Collected for the National Water-Quality Assessment Program, 1992-97, Water Resources Investigation Report 01-4178. NAWQA, Indianapolis, IN, 84 pages,**

**http://water.usgs.gov/pubs/immediate_release.html). This document is mostly discussing precision in the reproducibility concept, since it is discussing several years of data and dates and lab operators have changed.**

**Precision can be measured or estimated not only for physical or chemical measurements, but also for biological. For example, the State of Wyoming is discussing precision under reproducibility conditions when it defines:**

> **Precision as "Multiple duplicate samples at one site taken by samplers not communicating with each other; multiple sub-samples in adjacent reaches at a site having similar habitat and stressors; for visual based qualitative habitat assessments, two or more samplers doing independent site evaluations" (at http://deq.state.wy.us/wqd/watershed/10574-doc.pdf).**

**V-D. Field Measurement Precision:**

**For all measurements done in the field, Section V of the detailed study plan should briefly summarize how QC performance standards and estimation methods for field measurement precision were selected (for example, to be consistent with a listed State or Federal regulatory program). A statement in the study plan should explain that measurement quality objective and estimation details on this QC topic may be found in the applicable QC SOP included in each protocol.**

**For aquatic monitoring projects, the following guidance should be helpful in guiding what goes into the QC SOP under the heading of field measurement precision:**

## Typical Park Service QUALITATIVE DQOs:

**A typical minimum NPS default QC objective is that measurement precision under repeatability (or reproducibility, if more applicable) conditions should be controlled in a least one place, either the field or lab, depending on where the measuring is being done. Unless otherwise specified, all references to precision in this section refer to precision under repeatability conditions. When practicable, combined precision factors influenced by both lab and field factors should be considered together in a single step. A good default way to do this, when not required to quantify precision in other ways in defined regulatory methods, is to use the basic NAWQA model (http://water.wr.usgs.gov/pnsp/pest.rep/sw-t.html#QA) recommended by USGS:**

> **Sample replicates are designed to provide information needed to (1) estimate the precision of concentration values determined from the combined sample-processing and analytical scheme and (2) evaluate the consistency of identifying target analytes for the contaminant of concern. Each replicate sample is an aliquot of native sample water from a splitter and is processed immediately after the primary cone-split sample using the same equipment;**

> placed into the same type of bottle; prepared in the same way…if applicable; and stored and shipped in the same way.

**However, this single-step version of precision may be beyond the means of some monitoring units, so at minimum either lab or field measurement precision should be controlled.**

**The remainder of this section applies only to field measurement precision:**

## Typical NPS QUANTITATIVE Field-Precision Measurement Quality Objectives (QC Performance Standards):

**The plan should specify the frequency and percentage (of total samples) of QC samples to be required.**

**For parameters or estimates to be made in the field, at least one QC precision (either repeatability or reproducibility) sample should be analyzed in the field prior to analyses of other field samples that day at that site. Another QC sample should be taken if more than 20 samples are taken at that location. When measuring in the field, one is changing batches or conditions each day and/or whenever moving to a new location and may be impacting instrument calibration. Therefore, at least one QC sample should be taken at each site and/or each day, even if 20 samples are not taken at that site. Notes:**

> **This recommendation is also consistent with the commonly suggested advice pertaining to water column samples and to general samples that QC samples be done for each sampling set (batch), or at a rate of not less than 5%, whichever is more frequent (American Public Health Association, American Water Works Association, and Water Environment Federation. 1998. The 20th Ed. Of Standard Methods for the examination of water and wastewater, 20th Ed. American Public Health Association, Washington, D.C**
>
> **This is consistent with EPA guidance to take at least one QC sample in the field prior to field measurements and to have at least one set of QC samples per each 20 samples (5%, http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf**
>
> **In cases where data is to be compared with State data, check with State to make sure QC samples and frequencies are adequate.**
>
> **One often makes a couple of duplicate measurements of one sample to check repeatability precision at a site to make sure the instrument has settled down and thus two or more consecutive measurements agree or almost agree. The results of the final two measurements (of exactly the same sample) could be**

**recorded as precision repeatability duplicate measures. This would typically be done before the next step of measuring a known standard to check for systematic error. If one simply moves down the river in short order and measures another site, recalibration and/or another QC sample may not be necessary until one has reached the 20 sample limit. However, if one next put the electronic meters in a pickup and bounces down a rough road and goes to a site where chemical and physical (temperature, elevation, pressure, etc.) may be different, one has essentially started an new sample batch, and field calibration and new QC samples must be taken.**

**Field measurement repeatability precision for parameters measured by probes are typically obtained by duplicate (or more frequent) field measurement of the exactly the same homogeneous material at least once for each sampling trip. Unless otherwise justified, QC measurement quality objectives (MQOs), should not exceed a relative percent difference (RPD, for sample size of two –duplicates) or a relative standard deviation (RSD, for sample size of three or above) of 10%. Field probe measurements appropriate for this performance standard include pH, temperature, DO, specific conductance, salinity depth, light transmittance (PAR), turbidity, and Secchi Depth. Although these measurement quality objectives were designed for the marine or estuarine environments (Table A7-1 of the EPA QAPP at http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf) these QC performance standards are not especially stringent, so unless otherwise justified, MQOs at least this stringent should be used for freshwater samples too.**

**Some other parameters sometimes measured in the field have less stringent performance standards. For example, E-EMAP suggests that measurement precision for nutrients, chlorophyll a, and TSS not exceed 30%. Performance standards and frequency of QC samples should be those suggested by EPA EMAP for standard water quality probe parameters unless otherwise justified (http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf). Although RPDs are sometimes reported for a sample size as small as two, two is not a large enough sample size to give a reliable estimate of precision, and it should be kept in mind that 8-20 replicates, perhaps by pooling data over time, is necessary to keep measurement uncertainty down to reasonable levels (see step IV-F.2).**

**Those standardizing with USGS methods may alternatively use any more stringent performance standards or goals for field probe measurements recommended in the USGS National Field Sampling Manual for WRD. For example, for specific conductance, the USGS Field Manual specifies the following: for specific conductance values ≤ 100 uS/cm the repeatability (precision) goal specifies that repeat measurement should be within ± 5 % (RPD units?--manual doesn't give units) of each other, based on full scale measurements. For specific conductance values of > 100 uS/cm the repeatability (precision) goal) is for repeat measurements (an estimate of repeatability precision) to be within ± 3 percent of each other, based on RPDs and full scale measurements (http://water.usgs.gov/owq/FieldManual). This seems to be a case where lower (more stringent) goals are recommended than**

**the 10% precision performance standard EMAP recommends for field probe measurements in estuarine or marine environments.**

**However, if higher percentages than those recommended by EMAP are chosen, a justification should be given in the plan as to why such higher limits are necessary, adequate, and appropriate.**

**For NPS consistency, in all cases express precision achieved (in the case of either repeatability or reproducibility) in standard deviation units as suggested by NIST for uncertainty analyses and as is explained in step IV-F.2. If either RPD or RSD units are also to be reported, say to compare QC results with the performance standards of other programs, one typically uses the units suggested in those programs, often RPD units for sample size of two and RSD units for sample size of three or greater unless otherwise justified.**

**Notes on RSD vs. RPD units:**

**Why might some recommend RSD units as a first (default) choice default rather than recommending relative percent difference units? The answer is that one can always calculate RSDs, but can only calculate RPDs when sample size is two. Some also complain that RPDs can go no higher than 200%. Others would counter that RPDs are traditionally done for duplicates, and that 200% would not pass any reasonable precision performance standard anyway. Be aware that RPD and RSD values are typically different (for example an RPD of 20%, for a given set of numbers may be equal to relative standard deviation (CV units, where CV is the standard deviation divided by the mean, then multiply the result times 100) of 13%.**

**A data-comparability potential justification for using RPDs when sample size is two is that so many state and federal agencies, and labs in general by common convention use Relative Percent Difference (RPD) QC performance standards for precision rather than RSDs. Another potential justification is that standard deviations are not very well estimated when sample size is only two. Furthermore, for a given comparison of two, RSD's tend to be smaller than RPDs, so an RPD performance standard of 10% would be harder to meet for RPDs than for RSDs. For example, if the duplicates were 100 and 110, the RPD would be 9.5% and the RSD would be 6.7%. In this case the RPD would be closer to flunking the standard. It could be argued that this harder standard relative to RPDs is appropriate when sample size is only two (and therefore standard deviation estimates are very crude).**

**What about expressing precision as a plus or minus confidence interval, for example a 95% confidence interval? This is a more common process when one is expressing**

**uncertainty in a mean rather than expressing precision, but is not a bad idea as an additional but not required step. It is the seeming default in EPA's new STORET database. However, only report confidence intervals when sample size is 10 or more. The new STORET has a place for giving a confidence interval associated with precision QC results. It is not clear from EPA help fields, but we assume EPA is talking about a parametric t test confidence interval, such as the default confidence interval calculated as a default for smaller sample sizes in MS Excel, the NPS standard spreadsheet. Again for emphasis, such a confidence interval for precision should not be calculated and recorded in STORET if sample size is less than 10 (some would say 10 duplicates), a common scenario. If sample size is questionably too small to give a reasonably good estimate of the confidence interval and/or the QC precision results data gives any strong hints of non-normal distribution, a nonparametric confidence interval should be calculated and metadata about what was done and why should be included in the results comments field of STORET screen R4. Also remember that a true confidence interval on the mean reflects uncertainty in the mean precision as either repeatability (nothing changed) or reproducibility (something changed) and that one should specify which type of precision is being reported.**

**Field measurement precision of biological parameters or metrics, typically obtained by duplicate field measurement of the exactly the same homogeneous material, or in the case of destructive sampling a nearby spot in the same strata, should not exceed a stated performance standard. The best units to use are standard deviation units as suggested by NIST (see section IV-F.2 for details). Historically, and even today, however, the most common units used have been relative percent differences (RPDs, for sample size of two) or relative standard deviations (RSDs, for sample size of three or greater) precision performance standard. If the material analyzed is not exactly the same, the RSD (or relative percent differences if sample size is two) of metric differences are typically higher than for chemical analyses, perhaps 30% or more. For more details, see**

1. **Barbour, M.T., J. Gerritsen, B.D. Snyder, and J.B. Stribling. 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, Second Edition. EPA 841-B-99-002. U.S. Environmental Protection Agency; Office of Water; Washington, D.C. (http://www.epa.gov/owow/monitoring/rbp/ch04main.html#Section 4.4).**
2. **J.B. Stribling. 2003. Determining the quality of taxonomic data, J. N. Am. Benthol. Soc.22(4):621–631.**

**Some states define QC performance criteria for precision in biological observations (for example, see Wyoming criteria at http://deq.state.wy.us/wqd/watershed/10574-doc.pdf). State of MD QC performance standards (measurement quality objectives) for both biological values (like IBIs) and field measures like pH and nutrients are at http://www.dnr.state.md.us/streams/pubs/ea03-1qaqc.pdf/.**

**For regulatory monitoring, the QC performance standards may be used as "data acceptance" criteria. In this scenario, if the stated acceptable measurement quality objective performance standard is 20% and precision exceeds an RPD (sample size of two) or an RSD (sample size of three or more) of 20%, the data is typically not be accepted or reported, and if the monitoring staff gets this result, attempts would be made to increase measurement precision until the performance standard can be achieved.**

**Like EPA, the DOD recommends the use of data acceptance criteria based on QC results: "All quality control measures shall be assessed and evaluated on an ongoing basis, and quality control acceptance criteria shall be used to determine the usability of the data" (DOD Quality Systems Manual – Version 1 Final Based On NELAP Voted Revision 12 – 1 July 1999 (https://www.denix.osd.mil/denix/Public/Library/Compliance/EDQW/QSM.pdf).**

**.Alternatively, for general status and trends monitoring projects, especially those where monitoring programs are using USGS methods, the monitoring network may choose to flag those data having QC field precision measurements exceeding the QC performance standard. If USGS protocols are used, the data may reported with an flag signifying precision performance standards were not met and later interpreted USGS-style (taking precision into account) even if the QC performance standards are not met.**

**The Department of Defense takes the position that flagging of data due to problems in meeting QC acceptance criteria should only be used as a last resort, and that it is typically better to recalibrate, find and correct problems, etc. The data can be reported to meet the acceptance criteria (https://www.denix.osd.mil/denix/Public/Library/Compliance/EDQW/QSMver2.PDF).**

**In all cases, the QC performance standards to be used should be specified in the plan, and QC results should be specified two ways (as raw data and as summary statistics such as sample standard deviations, RPDs, RSDs, or confidence intervals) in metadata. In other words, if the two measures of the identical substance results in observations of 100 and 110, report 100 and 110 as part of QC metadata results for all data in that sampling trip, in addition to the summary statistics (such as an RPD or RSD) calculated for these two values.**

> **Note on Precision Metadata in the New STORET database: As explained in Part E (Figure 7, http://science.nature.nps.gov/im/monitor/protocols/wqPartE.doc) of this guidance, as regards chemical measurements, STORET relies mainly on a plus or minus field for "precision" in the CHEMICAL DATA RESULT ENTRY BOX. Since it is not clear what this means in terms of all potential types of field measurement precision samples, it is important to list all other important metadata related to precision in the Comments Box that is part of**

**the R4 chemical result data result entry box. For example, in addition to information in the plus or minus precision data entry box, how was the "plus or minus" figure calculated, and what were the actual QC sample results? Again, since new STORET suggests that precision be expressed in plus or minus units, for NPS consistency, the plus or minus should be expressed as plus or minus RPD percentage units unless otherwise justified. If the results of duplicate measures were 100 and 110, those results should be listed as 10% RPD if the (plus or minus) 10% was calculated as a relative percent difference. If 10% was calculated as a percent difference it should be given as 10% PD. If the plus or minus was calculated as a relative standard deviation, enter it as 10%RSD. If the field precision result is a consequence of a physical or biological measure rather than a chemical measure, then the actual results of the QC measures, and detailed metadata about the summary statistics, is not to be entered into a STORET CHEMICAL RESULT DATA ENTRY box, but elsewhere in STORET.**

**More Detail: Why is it so important to list the actual results of QC calculations related to precision, systematic error (bias), and overall uncertainty in accuracy? Because precision, systematic error (bias), and accuracy summary statistics are often given in different units, making it hard to bound minimum error as variability or minimum uncertainty (see step IV-F.2). Consider the following example. In the example where QC precision duplicate results were the two values were 100 and 110, the summary statistics that might be reported by various groups in any of the following ways:**

> **A sample variance of 50.**
>
> **A percent difference of 10%, or**
>
> **A precision duplicate relative percent difference (RPD) of 9.5% (the preferred alternative if sample size is two, or**
>
> **A precision duplicate relative standard deviation (RSD) of 6.7% (this is the NPS default preferred alternative if sample size exceeds two), or**
>
> **A standard deviation of 7.1, or**
>
> **For data obviously not normally distributed, and sample size of 6 or more, some might enter a precision duplicate nonparametric RSD (derived from f-pseudosigma and the median) of 3.5%.**
>
> **A 95% (t distribution) confidence interval of (a plus or minus) 63.5 in original units or a plus or minus 61% in percentage units about a mean of 105 (41.5 to 168.5) could be calculated from these two numbers (100 and 110), though it should not be in this case, since**

> **sample size is only two. A t distribution 95% confidence interval is the NPS preferred alternative is sample size is 10 or greater and the distribution of values shows no strong hints of non-normality. If a sample size of ten or more shows notable hints of non-normality, then a nonparametric 95% confidence interval is the preferred summary statistic for precision. Confidence intervals are the preferred ways to express precision in the new STORET database of EPA and are also recommended in the 20th edition of Standard Methods. However, confidence intervals should not be calculated when sample size is less than 10 (in this case, sample size is 2), a fact that neither Standard Methods nor STORET currently mention.**

**The above-listed options for summary statistics are not consistent (not expressed in comparable units). Therefore, data users need the QC result observation values themselves (i.e. 100 and 110) in order to enable others trying to use the data to convert all obvious contributors to measurement error or lack of confidence into comparable units. This is necessary before one can attempt to bound total measurement or study error or uncertainty (lack of confidence).**

**For habitat observation variables, the actual or true value can never be known with complete confidence, and thus the exact uncertainty in accuracy may be difficult to determine. In these situations, field measurement precision QC becomes an issue that can be controlled to a greater degree and thus becomes very important (EPA 1999. Aquatic Habitat Indicators and their Application to Water Quality Objectives within the Clean Water Act, EPA 910-R-99-014, Section 7, available at http://yosemite.epa.gov/R10/ecocomm.nsf/6da048b9966d22518825662d00729a35/f25bad58f59599058825679a005c6983/$FILE/Ahi_fina.pdf).**

**Monitoring experts should be aware that what field monitoring probe makers typically give as factory (advertisement) precision or "repeatability" specifications may not be achievable in real-world outdoor scenarios. Often, what they are calling precision or repeatability is different than what chemical laboratories mean by precision. However, the performance standards suggested by E-EMAP (http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf) as "maximum allowable precision goals", typically 10% RPD units is sample size is two or 19% RSD (if sample size is three or more) for field probe measurements of standard parameters like DO, pH, and temperature, are not all that stringent and should be achievable even under difficult field conditions. If not achieved, recalibration and/or technician retraining should be implemented.**

**For additional details on this topic and others related to field measurement precision, see appendix V-D.**

## V-E. Lab Measurement Precision:

**For all measurements done in the lab, Section V of the detailed monitoring plan should briefly summarize how QC performance standards and estimation methods for lab measurement precision were selected (for example, to be consistent with a listed State or Federal regulatory program). A statement in the study plan should explain that measurement quality objective and estimation details on this QC topic may be found in the applicable QC SOP included in each protocol**

**For aquatic monitoring projects, the following guidance should be helpful in guiding what goes into the QC SOP under the heading of lab measurement precision.**

**Laboratory measurement precision is the closeness of agreement (measured as variability) of repeat (or replicated for reproducibility) laboratory measurements. Measurements should be independent test results obtained under stipulated conditions, either repeatability (nothing in the measurement process changed) or reproducibility (something changed). In the repeatability context, it is a measure of the variability in results when one is measuring the same (homogenous) thing repeatedly in the laboratory, using the same operator, measuring device, under same atmospheric conditions, etc. Unless otherwise specified, all references to precision in this section refer to precision under repeatability conditions**.

## Typical Park Service QUALITATIVE Data Quality Objectives:

**A typical minimum NPS default QC objective is that measurement precision should be controlled in a least one place, either the field or lab, depending on where the measuring is being done. When practicable, combined precision factors influenced by both lab and field steps should be considered together in a single step. However, as discussed above, this single-step version of precision may be beyond the means of some monitoring units, so at minimum either lab or field measurement precision should be controlled. The remainder of this section applies only to lab measurement precision:**

## Typical NPS QUANTITATIVE Lab-Precision Measurement Quality Objectives (QC Performance Standards):

**Unless otherwise justified, Performance standards QC samples should be at least as stringent as those suggested for EPA estuarine and marine EMAP parameters (http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf).**

**Although the marine EMAP QAPP specifies inter-laboratory checks for differences in precision and systematic error, it evidently does not specify performance standards for precision reproducibility between labs. If wanted to insure that the use of multiple labs was not worsening precision, one could specify that the**

**performance standard for precision repeatability for one lab also be used as a performance standard for inter-lab precision reproducibility. In the case of the EPA marine EMAP, a typical precision repeatability performance standard for many lab measured water column parameters (such as nutrients, TSS, and Chlorophyll *a*) is a relative percent difference on no greater than 30%. If this same value was used for precision reproducibility between labs, maximum imprecision could be held to that level even if multiple labs were used.**

**Some might argue that the EPA marine performance standards are not especially stringent, but they are standardized, and they should be achievable. Where possible, especially for freshwater work, or when trying achieve data comparability with other data sets that use more stringent precision QC measurement quality objectives, more stringent standards should be used**

**The plan should specify the frequency or number of QC samples.**

**For lab work to be done on marine or estuarine samples, frequency of QC samples should be at least as stringent as those suggested for EPA estuarine and marine EMAP parameters. Commonly, EMAP requires analytical sets or batches should be held to 20 or less samples and must include appropriate QC samples uniquely indexed to the sample batch. The number of QC samples required per batch differs according to the analyte (http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf).**

**For toxic chemicals analyzed in fish tissues or other solid matrices such as soil or sediments, unless otherwise justified, use quality control least as rigorous as that suggested by EPA or data sets the new data. A frequency of one QC sample per 20 sample batch (5%) is again suggested, except that for "calibration check standards,", two QC samples (10%) are advised  (EPA, 2000, Guidance for assessing chemical contaminant data for use in fish advisories, Volume 1: Fish Sampling and Analysis - Third Edition. EPA 823-B-00-007, Table 8-7, available on the Internet at http://www.epa.gov/ost/fishadvice/volume1/v1ch8.pdf).**

**This recommendation is also consistent with the commonly suggested advice for water column and other samples that QC samples be done for each sampling set (batch), or at a rate of not less than 5%, whichever is more frequent (American Public Health Association, American Water Works Association, and Water Environment Federation.  1998.  The 20th Ed. Of Standard Methods for the examination of water and wastewater, 20th Ed.  American Public Health Association, Washington, D.C.).**

**Alternatively, for freshwater samples, in cases where data is to be compared with USGS data, the number of QC samples to be analyzed in the lab compared to total samples should be those suggested by USGS. For example NAWQA guidance for QC design call for at least 1 QC sample per 20 samples, but some types of QC samples and analytes call for different specifications. There are some differences between field blanks, replicate field matrix spikes, and replicates (USGS**

**terminology for precision repeatability samples). For details see Table 1 at http://water.usgs.gov/nawqa/protocols/OFR97-223/ofr97-223.pdf. USGS QC methods and terminology are different enough that if USGS data comparability is critical, it might be easier to arrange for USGS to perform all the sampling, lab analyses, and QA/QC.**

**Since new STORET suggests that precision be expressed in plus or minus units, for NPS consistency, the plus or minus should be expressed as plus or minus RPD units for sample size of two or RSD percentage units (for sample size of three or more) on form R4 unless otherwise justified.**

**New STORET has a place for giving a confidence interval associated with precision QC results. It is not clear from EPA help fields, but we assume EPA is talking about a parametric t test confidence interval, such as the default confidence interval calculated as a default for smaller sample sizes in MS Excel, the NPS standard spreadsheet. Such a confidence interval should not be calculated and recorded in STORET if sample size is less than 10, a common scenario. If sample size is questionably too small to give a reasonably good estimate of the confidence interval and/or the QC precision results data gives any strong hints of non-normal distribution, the confidence interval box should be left blank or a nonparametric confidence interval should be calculated and metadata about what was done and why should be included in the results comments field of screen R4.**

**When not specified by the SOP, the general default for nutrients, organics, other chemicals, and biological measures, lab measurement precision should not exceed a relative percent difference of 20% when sample size is two or a RSD or 20% when sample size is greater than 2.**

**Matrix spike duplicates and sample duplicates are sometimes more difficult for certain analytes. For example, DOD recommends a QC acceptance criteria of less than or equal to 30% RPD for many GC or HPLC organic lab analyses and for some metals analyses, and less than or equal to 20% for ICP metals (Tables B-1, and B-5, https://www.denix.osd.mil/denix/Public/Library/Compliance/EDQW/QSMver2.PDF).**

**Limitations that specify different performance standards depending on the closeness of the measured value to the instrument detection limit may also be specified. For example, a superfund performance standard for ICP analysis for metals states “A control limit of plus or minus 20% for the Relative Percent Difference (RPD) should be used for original and duplicate sample values greater than or equal to 5x the Superfund Contract Lab Program contract required detection limit (CRDL)…A control limit of plus or minus the CRDL should be used if either the sample or duplicate value is less than 5x CRDL. In the case where only one result is above the 5x the CRDL level and the other is below, the plus or minus the CRDL criteria applies. If both samples values are less than the IDL (sic, MDL or 3 x the SD), the**

**RPD is not calculated (http://www.epa.gov/r10earth/offices/oea/fginorg.pdf). An EPA region 4 lab requires the same basic limit:**

> **Matrix Duplicate Samples: At least one matrix duplicate sample will be prepared for each batch of samples (or perhaps lab run, whichever is most applicable). Evaluate the matrix duplicate sample against the following criteria: For those results greater than 5X the MQL results should be within 20% RPD. For those analytical results less than 5X the MQL, duplicate results should be within plus or minus the MQL of each other. If one result is >5X the MQL and the other is <5X the MQL, results should be within + the MQL of each other. If duplicate sample results do not meet these criteria and the method has been demonstrated to be in control with the LCS, flag the analyte in the duplicate sample as estimated (J) and add the remark: "Matrix precision outside method control limits for _____" (http://www.epa.gov/region04/sesd/asbsop/asb-loqam.pdf).**
>
> **Note: For regulatory monitoring, the QC performance standards may be required by a state or federal agency to be used as "data acceptance" criteria. In this scenario, if the precision exceeds the QC performance standard, the data should not be accepted or reported. If the monitoring staff gets this result, attempts would be made to increase measurement precision until the performance standard could be achieved, and data should not be reported until it meets the performance standard.**

**If USGS protocols are used, USGS-style metadata should be used.**

> **Note: For general status and trends monitoring projects, especially those where monitoring programs are using USGS methods, the monitoring network may choose to flag those data having QC field precision measurements exceeding the QC performance standard. Data reported with such a flag notation could then be interpreted USGS-style (taking precision into account) even if the QC performance standards are not met.**

**In all cases, the QC performance standards to be used should be specified in the plan, and QC results should be specified two ways (as raw data and as summary statistics) in metadata. In other words, if the two measures of the identical substance results in observations of 100 and 110, report 100 and 110 as part of QC metadata results for all data for that trip, in addition to the summary statistics (such as a RPD) calculated for these two values.**

> **Note on Lab Precision Metadata in the New STORET database: As explained in Part E (Figure 7) of this guidance, as regards chemical measurements, STORET relies mainly on a plus or minus field for "precision" in the CHEMICAL DATA RESULT ENTRY BOX. Since it is currently not totally clear what this means in terms of all potential types of lab measurement precision samples, it is important to list all other important metadata related to precision in the Comments Box that is part of the R4**

chemical result data result entry box. For example, in addition to information in the plus or minus precision data entry box, how was the "plus or minus" figure calculated, and what were the actual QC sample results?

Again, if the results of duplicate measures were 100 and 110, those results should be listed as raw QC results. In the plus or minus box for precision, enter 9.5% RPD units, since sample size is only two. A confidence interval should not be calculated if sample size is less than 10, as in this case. Again, since new STORET suggests that precision be expressed in plus or minus units, for NPS consistency, use RPD units for sample size of two or RSD units for larger sample sizes unless otherwise justified.

**For more details on this step, see appendix V-E.**

## Introduction to Systematic Error (Bias):

**Measurement systematic error (bias) is present if measurement results are consistently either too high or too low in comparison with the right answer (or the expected or correct answer). Such errors are typically in one direction only (either too high or too low, either positive nor negative), and they are said to be "systematic errors."**

**Many use the phrase "systematic error" as a synonym for systematic error (bias). NIST prefers the use of systematic error, stating that some restrict the word bias to instruments only (http://physics.nist.gov/Document/tn1297.pdf). Herein, we follow this same common and consensus understanding and use the two terms as synonyms, but also tend to use the phrase "systematic error (bias)" to make it clear that the two concepts are the same in our discussions. However, an important thought to grasp is that it is typically necessary to distinguish between average amounts of systematic error, herein referred to as systematic error (bias), expressed as a mean of bias estimates (usually one for more two-number expected vs. observed value comparisons), and the component of uncertainty arising from a systematic effect. The later is typically expressed as a variance and added other variances (such as the variance for precision) in sum of squares calculations of expanded combined standard uncertainty (see further explanation in step IV-F.2, uncertainty analysis). For now just remember that systematic error (bias) is typically one's best estimate of an average amount of systematic error (bias) from two or more estimates, so it is typically a mean. On the other hand, "the component of uncertainty arising from a systematic error (bias)" (systematic effect) is typically the sample variance of those same values used to determine the mean. This component of uncertainty as a variance is typically added to root sum of squares (RSS) uncertainty equations only when the raw data has not been corrected or adjusted for average amounts measurement uncertainty (both bias and imprecision), the regrettable but common situation in many contemporary water quality or contaminants studies.**

**Common potential sources of chemical measurement systematic error (bias) include recovery errors and blank control systematic error (bias).**

**Recovery errors occur when one recovers either too little (say 80%) or too much (say 120%) versus the right answer, which would be 100% when trying to recover a known value (100%) from a certified reference sample (known most probable value) or a theoretical expected value (100%) of the analyte known to have been added to a matrix spike.**

**Sources of systematic error (bias) can include calibration errors, unaccounted-for interferences, or chronic sample contamination (as typically controlled with blanks). The sample itself may generate real or apparent systematic error (bias) caused by a matrix effect or variation in physical properties (http://www.epa.gov/quality/qs-docs/g5-final.pdf).**

**Another typical source of negative systematic error (bias) is sample degradation, such as happens when one exceeds maximum holding times. In one example, the maximum holding time for pH and dissolved oxygen samples is zero (analyze immediately, 40 CFR Part 136.3, see Web at http://www.access.gpo.gov/nara/cfr/cfrhtml_00/Title_40/40cfr136_00.html). In other words, measure these parameters in-situ in the field rather than carting a sample off the lab and bringing up the likely possibility that the pH or dissolved oxygen will change while transporting the sample to the lab.**

**Systematic error (bias) is often (wrongly) considered a synonym for accuracy, or confused with accuracy, depending on the author. Some try to distinguish between biases as a difference between the measured value and the expected value in matrix spikes vs. "accuracy" (sic) as the difference between the measured value and a known certified reference material value. In accordance with the recommendations of NIST and ISO, it is suggested that to avoid confusion all NPS documents use the phrase "systematic error (bias)" for the concept of measurement bias, and limit the word accuracy to discussions of uncertainty that include (at minimum) systematic error (bias) and precision under conditions described as either repeatability or reproducibility.**

**ISO defines systematic error (bias) as the difference between the EXPECTATION of the test results and an accepted reference, which to many investigators sounds a bit like accuracy. Clearly however, overall (uncertainty in) accuracy is influenced by both measurement precision and measurement systematic error (bias). Since systematic error (bias) is only one part of the calculation of overall uncertainty in accuracy, bias cannot logically be considered a synonym for either accuracy or uncertainty. Lab measurement systematic error (bias) is simply producing results that are consistently too high to too low.**

**No matter whose definition of systematic error (bias) one looks at, a common thread is that one needs to know what the right, or expected answer is, then measure**

**something with that known value, then record the average systematic error (average bias) as the difference between the average value obtained vs. the "right" at least the "expected" answer.**

**In many toxic chemical analyses, certified reference materials have a known or expected value and are used to determine systematic error (bias). Labs usually report this as percent recovery.**

**In biological and some other types of measures or observations, things are often more challenging and sometimes the observation of an expert is considered "right" and the difference between that observation and those of rookies or trainees is considered systematic error (bias). Other parameters that don't lend themselves to easy estimates of systematic error (bias), include:**

10. **PAR (light attenuation in lakes or marine environments)**
11. **Bacterial Counts (fecal coliforms, E. coli, etc.)**
12. **Taxonomic Identification of Very Small Invertebrates or Other Difficult Taxa**
13. **Judgment Habitat Observations (Percent Embededness of Cobbles)**
14. **Spike Recoveries of Chemicals in Difficult Matrices**
15. **Dissolved Oxygen Concentrations**
16. **5-day Biological Oxygen Demand**
17. **Field Measured Temperatures**
18. **Laser and other New Technologies That Seem to Measure Things Better than Old Technologies (so which answer is most accurate/has the least uncertainty in accuracy?).**

**In these types of parameters, the expected or correct answer is not always easy to identify, especially if one is doing the measurement in the field, so systematic error (bias) is more difficult to estimate and it might be more tempting than usual to confine uncertainty estimates to precision as either reproducibility or repeatability aspects or even to (relatively crude) rounding rules (see discussion of rounding rules and how they relate to uncertainty in accuracy in section V-I and appendices). However, even for these types of parameters, there is often some way to at least roughly estimate systematic error (bias). For example, sometimes the observation or estimate of an expert is considered "right" or "expected" result and the difference between that observation and those of rookies or trainees is considered systematic error (bias). In cases where one is not sure even an expert is "right", another approach would be to take the maximum difference (delta) between observations as systematic error (bias). In this case, the systematic error (bias) estimate would be the same as maximum difference in reproducibility (something changes) precision. That would then be the conservative (worst case) estimate of bias and the variance of that value would be added to the variance of the value for precision repeatability (nothing in the measurement process changes) in sum of squares NIST calculations of measurement uncertainty. This approach would perhaps overestimate systematic error (bias) and express it as a plus or minus factor, perhaps not a bad thing when**

**the right answer is not easy to pin down. An approach such as this one would allow one to estimate NIST measurement uncertainty, and in most cases would still be superior to trying to use rounding rules as crude ways to account for uncertainty.**

**For biological taxa species identification and enumeration, EMAP indicators guidance (http://www.epa.gov/emap/html/pubs/docs/resdocs/ecol_ind.pdf) recommends:**

**· 10% of all samples are checked for accuracy [sic, they mean systematic error (bias) when they say accuracy] by a senior taxonomist.**
**· Re-identification samples are randomly chosen (1 out of 10) on a regular basis (this would be precision as repeatability if done by the same person or precision as reproducibility if done by two individuals or labs).**
**· Accuracy (sic, bias/systematic error) will be calculated as:**
**Total # of organisms in QC recount -Total # of errors x 100**
**Total # of organisms in QC recount where errors include:**
**Counting error (e.g., counting 11 of a given species instead of 10)**
**Identification error (e.g., misidentifying species X as species Y)**
**Unrecorded taxa errors (e.g., not identifying species X when it is present)**
**· Actions for unsatisfactory level of taxonomic accuracy:**
**90-95% - Original technician advised, species identifications reviewed, and any changes to species**
**Identifications recorded on original data sheet.**
**< 90% - Same as for 90-95% but numerical counts should also be corrected on original data sheet.**

**Also, for “accuracy” and consistency of field identifications among field crews, EMAP suggests the following QC measures:**

**• Consistent training**
**• Performance evaluations against experts**
**• Use of experienced personnel (Federal, state and university)**
**• Consistent protocol for vouchering specimens for confirmation of species identifications to allow for data correction when necessary**
**• Field audits**

**With biological collecting gear, the amount of systematic error (bias) might relate to the fact that certain collecting gear is biased only towards biological specimens of certain size and weight.**

**Some states define QC performance criteria for systematic error (bias, often wrongly referred to as “accuracy”) in biological observations. For example, Maryland QC performance standards (measurement quality objectives) for “accuracy” of biological values (like index of biological integrity (IBI) metrics at http://www.dnr.state.md.us/streams/pubs/ea03-1qaqc.pdf/. Wyoming Bioassessment guidance for systematic error includes systematic error (bias) samples such as taxonomic reference samples and spiked organism samples, as well as ”accuracy” (sic, they mean systematic error/bias) controls that involve the confirmation of**

**identifications and calculations of percentage of missed specimens (http://deq.state.wy.us/wqd/watershed/10574-doc.pdf),**

**Chemists are typically more accustomed to controlling for systematic error (bias) than biologists. Systematic error (bias) is a bit more straightforward in chemical analyses, especially when an NIST standard or NIST traceable standard can be used as certified reference material, since the "right" answer is known. Matrix spikes are a bit less straightforward. There one is looking for an "expected" answer.**

**NIST and ISO suggest that data should be corrected (adjusted) to make up for known systematic error (bias). Although the reasons for not doing so have not always been sound, historically, this has not been done in the water quality or contaminants monitoring in the US very often. For more details, see discussion about adjusting data for systematic error section IV-F.2 and appendix IV-F.2.**

**V-F. Lab Measurement Systematic Error (Bias):**

**Section V of the detailed study plan should briefly summarize how QC performance standards and estimation methods for lab measurement systematic error/bias were selected (for example, to be consistent with a listed State or Federal regulatory program). A statement in the study plan should explain that measurement quality objective and estimation details on this QC topic may be found in the applicable QC SOP included in each protocol.**

**For aquatic monitoring projects, the following guidance should be helpful in guiding what goes into the QC SOP under the heading of lab measurement systematic error.**

## Typical Park Service QUALITATIVE Data Quality Objectives For Lab Measurement Systematic Error (Bias):

**If water quality or aquatic contaminants monitoring is designed so that the data to be collected will be comparable to USGS data, or if no specific systematic error (bias) estimation method is required by a regulatory agency, the USGS method for systematic error (bias) control and bias metadata reporting should be used.**

> **Note: this is a good default option since it includes systematic error (bias) introduced both in sample collection and processing. A typical minimum NPS default QC objective is that measurement systematic error (bias) be controlled in a least one place, either the field or lab, depending on where the measuring is occurring. It is more ideal, though, when practicable, to consider combined systematic error (bias) factors from both the field and lab in estimating systematic error (bias) in a single step, NAWQA style (http://water.wr.usgs.gov/pnsp/pest.rep/sw-t.html#QA by using:**

> **Field-matrix spikes designed to (1) assess recoveries from field matrices and (2) assist in evaluating the precision of results for the range of target analytes in different matrices.**
>
> **A field-matrix spike is prepared by adding a standard spike solution …to a split of sample water processed in the same way as the regular contaminant analysis.**
>
> **The USGS National Water Quality Lab (NWQL) in Denver sells spikes for the many compounds, but also purchase them from other vendors. Non-USGS monitoring groups could go direct to Supelco or another manufacturer.**

**However, considering both lab and field systematic error (bias) influences on overall systematic error (bias) as one step may be beyond the means of some NPS monitoring groups.**

**Performance standards for lab measurement systematic error (bias) must be listed. Unless otherwise justified, performance standards and frequency of QC samples for lab systematic error (bias) control should be at least as stringent as those suggested for EPA estuarine EMAP parameters, such as a maximum of 10% percent deviation from the "true" value for water column parameters such as chlorophyll a, TSS, and nutrients (Table A7-1 of the EPA Marin EMAP QAPP at (http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf).**

**Alternatively, if needed for data comparability, performance standards could be those specified by another large (comparable data) federal sampling program such as those of NOAA, FWS, or USGS.**

**If other performance standards and frequencies do not seem applicable, those recommended by the EPA Superfund CLP program could be used. For example, for the CLP program, inorganic analyses, matrix spike recoveries of 75-125% in solids are typically acceptable. For organic analysis, matrix spike recoveries are variable depending on the type of compound and matrix. CLP Recoveries of 70-130% are used as advisory only if sufficient recovery data are not available http://www.epa.gov/oerrpage/superfund/programs/clp/guidance.htm).**

**Alternatively, for freshwater samples, the number of QC lab samples compared to total samples could be those suggested in a document on QC design in NAWQA (http://water.usgs.gov/nawqa/protocols/OFR97-223/ofr97-223.pdf).**

> **STORET Note: QC sample results related to systematic error (bias) and accuracy are assigned to "trips" (individual sampling events). As explained in Part E (see Figure 3 and related text) of this guidance, "a trip occurs, for example, when a data collector leaves the office and collects samples and/or makes measurements/observations at the six stations included in the park's monitoring network and then returns to the office. It is included as a way to**

**attach QC data." This might include a reagent (method) Blank or a lab equipment blank contributing to lab measurement systematic error (bias) to all the samples collected during the trip. QC results related to any of these types of samples should be presented, not just summary statistics. Since new STORET relies mainly on a plus or minus field and a confidence interval field in the CHEMICAL DATA RESULT ENTRY BOX related to precision only (not systematic error/bias), and since not all explanatory information on types of laboratory spikes and reference materials are listed as standard STORET choices, it is important to list all other important metadata related to systematic error (bias) and/or overall accuracy in the Comments Box that is part of the chemical data result entry box.**

> **Note: In new STORET, the window for QC Sample (QC6) is the area where one records the results from QC samples including recovery of certified reference materials (also called laboratory control standards) or spikes (as well as metadata about transportation and storage methods.**

**Care needs to be exercised when considering the use of default QA objectives originating in USGS WRD offices, to make sure the NPS user knows what USGS means by the objectives. Until such time as USGS finalizes their nationwide QAPP and make terms clear and/or more consistent with EPA, the States, and other laboratories, it might be easy to misunderstand the USGS summary tables. For example, one USGS Texas QAPP on the Internet (http://www.dfwstormwater.com/FY01/PDFs/appendix_3_USGS.pdf) (go to http://www.dfwstormwater.com then search for quality assurance, then choose 01appendix_3_usgs.pdf) seems to call for QA accuracy (sic, they mean bias/systematic error) objectives for nutrients to be "within a plus or minus 3 standard deviations of known standard concentrations." A standard deviation is calculated on two or more numbers, and it is not clear which numbers USGS is talking about. When most agencies limit systematic error (bias), they typically use "% recovery", so the numbers involved are typically the "% recovered" value AND the expected number (100%). To be consistent with common convention, one might be expecting USGS to be talking about those same two numbers when giving a standard deviation as accuracy QA objective. Thus, at first glance, the Texas USGS table calling for accuracy of plus or minus 3 standard deviations (100% of the observations if the distribution was normal) would seem to be hinting that if the known certified reference material standard for nitrate was 100, and the lab got 200 as an answer (200% recovery) for one trial, the acceptable performance would be anything within 3 standard deviations (sample standard deviation of the two values 100 and 200 = 70.7, and 3 times that = 212.1). Since the known standard was 100, the acceptable performance would be 0 to 312% (100 plus 212) percent recovery? A recovery of 312% would not be acceptable to the NPS. However, in discussing this issue with USGS experts, it turns out that accepting any result within 3 standard deviations of the expected result is not what they meant. They were really talking about a different concept, the acceptability of the standard being used, and getting**

**answers within the same interval obtained by the supplier, though calibration solution suppliers use variable methods to determine the plus or minus "accuracy" (sic, they mean crude tolerance) intervals on the bottle labels. If the supplier of a standard labeled as "100 plus or minus 1" (the "right or expected" answer would be 100. If the supplier then informed the USGS that the standard deviation that the supplier had obtained for multiple samples of that particular standard was 4, then USGS staff would expect all samples analyzed by USGS from such standards to also fall within plus or minus 3 standard deviations (3 times 4), or 88 to 112. The USGS then uses that interval as control limits and considers anything beyond that limit as "out of control," something in the measurement process needing correction. So although USGS does not always reject or even annotate data when batch (or lab run) systematic error (bias) samples go beyond those limits, they do try to correct problems when beyond those limits. Thus, for all practical purposes, this plus or minus 12% control limit interval is roughly comparable to the EPA maximum allowable systematic error (bias) goal (acceptance performance standard) for nutrients in water, plus or minus 10%, as recommended by EPA's E-EMAP program http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf.
However, this already cloudy picture is further clouded by the fact that what most water monitoring specialists use for calibration standards do not specify a standard deviation representing NIST uncertainty, but rather specify plus or minus intervals that are crude manufacturing acceptance intervals rather than either NIST standard deviations or true statistical tolerance intervals.**

**In fact, the plus or minus accuracy interval on a calibration standard bottle, unless the standard is supplied by NIST, is not typically a standard deviation, even if the label says "NIST traceable." A check of several calibration standard suppliers for parameters like specific conductance and pH suggested that the interval instead was simply a crude manufacturing tolerance (sic, they really mean acceptance) interval, and not even a statistical tolerance interval. Often the manufacturer calibrates a measuring instrument with an NIST standard, and/or uses a highly accurate, three decimal place measuring instrument, then measures the sample batch to be distributed to make sure it is within the stated "accuracy" (sic, they mean acceptance) interval. If it is not, titrations are done to adjust the concentration of the batch until it is within the stated interval. So typically the calibration standard intervals provided are neither standard deviation intervals nor true tolerance intervals, but simply an acceptance interval tied to a standard operating procedure used in manufacturing calibration standards to be sold.**

**Lack of perfect purity in a reference material is simply one more source of uncertainty to combine with other uncertainty factors in calculating NIST uncertainty (see http://physics.nist.gov/Document/tn1297.pdf and discussion for step IV-F.2).**

**As recommended by EPA, when practicable, laboratory measurement "systematic error (bias) assessments should be based on analysis of matrix-spiked samples rather than reference materials so that the effect of the matrix on recovery is**

**incorporated into the assessment." When used, matrix spikes should be "added at different concentration levels"…For some measurement systems (e.g., continuous analyzers used to measure pollutants in ambient air), spiking samples may not be practical, so assessments should be made using appropriate blind reference materials" (http://www.epa.gov/quality/qs-docs/g5-final.pdf).**

> **Note: why are matrix-spiked samples important in chemical analyses? The answer is because different environmental matrices may have different interference characteristics and thus practical achievable recoveries may be different in one type of matrix vs. another. Though some worry most about screening methods, even so-called definitive methods are far from foolproof. Even methods such as ICP-AES and ICP-MS are not free from interferences that can compromise data quality. Practical interferences are valid in biomonitoring too. For example, great depth or current can interfere with the ability of a small seine to collect fish. For more details, see**
>
> > **J. Diamond et al. 2001. Towards a definition of performance-based laboratory methods. A position paper of the National Water Quality Monitoring Council Methods and Comparability Board, Technical Report 01-02, Web:** http://water.usgs.gov/wicp/acwi/monitoring/nwqmc), **and**
> >
> > **D. M. Crumbling. 2001. Applying the concept of effective data to Environmental analyses for contaminated sites, Current Perspectives in Site Remediation and Monitoring. EPA 542-R-01-013, www.clu-in.org, choose publications and studio, then Characterization and Monitoring).**
>
> **If one gets low (less than 60%) recoveries from spikes or certified reference materials, or if one routinely detects the analyte in laboratory blanks, consideration should be given to reporting values obtained as estimates, as some in USGS have done (D.W. Kolpin et al. 2002. Pharmaceuticals, Hormones, and Other Organic Wastewater Contaminants in U.S. Streams, 1999-2000: A National Reconnaissance, Environ. Sci. Technol. 36:1202-1211, http://pubs.acs.org/subscribe/journals/esthag/36/i06/es011055j.pdf).**

**When one finds a matrix spike result is consistently biased high or low, to avoid confusion, it is suggested that this type of QC comparison result be termed a quantification of "matrix spike measurement systematic error (bias)." On the other hand, if one finds a certified reference material result is consistently biased high or low, to avoid confusion, it is suggested that this type of QC comparison result be termed a quantification of "certified reference material lab measurement systematic error (bias)." It should also be made clear whether or not all operations were done in the lab, or whether the certified reference material (CRM, also called laboratory control standard) and/or spike went through field steps so that systematic error**

**(bias) components that are due to both sampling and analytical operation were included.**

## Method Specificity Control, A Concept Related to Systematic Error (Bias):

**Like interferences or lack of blank control, lack of measurement specificity control (QC control to ensure that the parameter of concern and not some other parameter is being measured) can influence systematic error (bias). For example, if a method is not specific enough, a false positive may occur when some other parameter is really the cause for the positive result. This would bias (M error) the results high. Although most EPA QA/QC do discuss related issues like blank control and interferences, most (including the EPA QA/QC documents G4, G5, and G5i at www.epa.gov/quality) such EPA documents do not address method specificity per se. However, method specificity control is a PBMS basic (J. Diamond et al. 2001. Towards a definition of performance-based laboratory methods. A position paper of the National Water Quality Monitoring Council Methods and Comparability Board, Technical Report 01-02, Web: http://water.usgs.gov/wicp/acwi/monitoring/nwqmc. Furthermore, in certain types of testing, such as DNA/RNA fingerprinting, method specificity can vary inversely with sensitivity (Peter Dratch, NPS, Personal Communication, 2002). In cases like these where method specificity is in doubt or otherwise especially important (as in DNA/RNA work), the team should consider specifying separate data or measurement quality objectives for method specificity.**

## Interferences Contributing to Systematic Error (Bias) in Habitat Observations, Biological Observations or Method-Defined Measures:

.

**Interferences that can impact systematic error (bias) (and therefore indirectly, overall uncertainty in accuracy) must also be controlled in biological studies, habitat observations, and method-defined measures, though the approaches can vary. Whereas in chemical studies, one can typically see if a method is giving a “right” answer by measuring a known reference material or a spike, the “right answer is not always as easy to come by in biological, habitat, or method defined measures (such as BOD, oil and grease, some bacterial counts, some biological observations, and many toxicity studies. However, as brought up first in our introduction to systematic error (bias, discussed above) one should not just automatically give up and not try to control for systematic error (bias) error in such cases. Here are some typical approaches:**

1) **Use defined reference conditions (clean, un-impacted reference sites) for biological assessments.**

2) **Use the results of a recognized expert as "the right answer" and then bound operator systematic error (bias) with the differences from the result of the expert vs. the trainees. For example, a true habitat expert could rate a site of "degree of embededness", percent cover, and other semi-quantitative habitat observations, and that expert's observations could be then contrasted to those of rookies to at least roughly approximate operator systematic error (bias).**

   **Additional discussion: In a similar manner, for lab-measured biological observations, such as lab taxonomic identifications of organisms, a slightly different approach is usually taken compared to the one listed below for chemical parameters. If a senior, recognized regional or national expert is identifying general macro-invertebrates, that expert's answers might be presumed to be correct. If a student or other non-expert were involved in identifying some of the samples, the identifications would be presumed to different, and less accurate, perhaps through operator systematic error (bias) in one direction or the other. A measure of systematic error (bias) then would be a comparison of the same-sample difference between the results of the rookie or trainee vs. the results of the expert. For additional discussion of systematic error (bias) in biological metrics, see**

   1. **Barbour, M.T., J. Gerritsen, B.D. Snyder, and J.B. Stribling. 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, Second Edition. EPA 841-B-99-002. U.S. Environmental Protection Agency; Office of Water; Washington, D.C. (http://www.epa.gov/owow/monitoring/rbp/).**
   2. **Additional ideas for systematic error (bias) (they call it accuracy) performance standards for biological observations may be found in Table 2-1 at http://www.sccwrp.org/regional/98bight/qaqc/qapln.htm#tab2-1 and in J.B. Stribling. 2003. Determining the quality of taxonomic data, J. N. Am. Benthol. Soc.22(4):621–631.**

3) **For method-defined measures, rigidly follow all SOPs and calibration recommendations and for things like BOD and oil and grease, consider clean-well aerated streams for systematic error (bias) control of blanks. It is also a good idea to do split samples and cross comparisons of your results with those of recognized expert labs. In such cases (and for the expert defined systematic error (bias) approximations discussed above for biological and habitat observations) where the systematic error (bias) control is not optimal, one should redouble efforts to control other PBMS or MQO measurement performance standards, such as those for precision, sensitivity, and performance.**

**For more details, see the papers from which most of information related to the three steps above was condensed:**

> **J. Diamond et al. 2001. Towards a definition of performance-based laboratory methods. A position paper of the National Water Quality Monitoring Council Methods and Comparability Board, Technical Report 01-02, Web:** http://water.usgs.gov/wicp/acwi/monitoring/nwqmc).
>
> **The EPA Internet summary comparing different Biomonitoring approaches based on a QC performance (PBMS) differences in interferences (as well as precision, systematic error (bias), sensitivity, and performance range) (Barbour, M.T., J. Gerritsen, B.D. Snyder, and J.B. Stribling. 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, Second Edition. EPA 841-B-99-002. U.S. Environmental Protection Agency; Office of Water; Washington, D.C., http://www.epa.gov/owow/monitoring/rbp/ch04main.html#Section 4.3).**

**The plan should document spiking protocols, and they should be explained in STORET.**

**The plan should specify that QC measurement systematic error (bias) performance for each measurement parameter should be reported in metadata and specified either as "certified reference material laboratory measurement systematic error (bias)" and/or as "matrix spike laboratory measurement systematic error (bias)." In STORET, choose reference sample in the Sample Data Entry box, and choose either Reference Sample or Field Spike. The actual QC performance results (say 100 and 110) should be included in STORET meta data, rather than just reporting a recovery of 110%). Use the comments field for any clarifications necessary in STORET.**

**Other regulatory or general status and trends monitoring default QC performance standards, rather than the ones listed below, may be used when appropriate to ensure data comparability. In all cases, the QC performance standards used should be specified in the plan, and QC results should be specified two ways (as raw data and as summary statistics) in metadata.**

**The rest of this section is devoted to the discussion of control of quantitative lab measurement systematic error (bias) alone.**

# Typical NPS QUANTITATIVE Measurement Quality Objectives (QC Performance

## Standards) for Lab Measurement Systematic Error (Bias):

**The plan should specify the frequency or number of QC samples.**

**For lab work to be done on marine or estuarine samples, performance standards and frequency of QC samples should be at least as stringent as those suggested for EPA estuarine and marine EMAP parameters. Commonly, EMAP requires analytical sets or batches should be held to 20 or less samples and must include appropriate QC samples uniquely indexed to the sample batch. The number of QC samples required per batch differs according to the analyte. As an example, the minimum QC samples required for nutrient analysis on a per batch basis include a four point standard curve for each nutrient of interest; reagent (method) blanks at the start and completion of a run; one duplicated sample; and one reference treatment for each nutrient (http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf).**

**For toxic chemicals analyzed in fish tissues or other solid matrices such as soil or sediments, unless otherwise justified use quality control least as rigorous as that suggested by EPA or data sets the new data. A frequency of one QC sample per 20 sample batch (5%) is again suggested, except that for "calibration check standards,", two QC samples (10%) are advised (EPA, 2000, Guidance for assessing chemical contaminant data for use in fish advisories, Volume 1: Fish Sampling and Analysis - Third Edition. EPA 823-B-00-007, Table 8-7, available on the Internet at http://www.epa.gov/ost/fishadvice/volume1/v1ch8.pdf).**

**This recommendation is also consistent with the commonly suggested advice for water column and other samples that QC samples be done for each sampling set (batch), or at a rate of not less than 5%, whichever is more frequent (American Public Health Association, American Water Works Association, and Water Environment Federation. 1998. The 20th Ed. Of Standard Methods for the examination of water and wastewater, 20th Ed. American Public Health Association, Washington, D.C.).**

**For metals analyses, lab measurement systematic error (bias), (typically quantified as % recovery of a known-value normalized to 100% and expressed as comparison % recovery) for a certified reference material sample, should not exceed a % recovery interval of 90-110%. A matrix spike comparison should not exceed a % recovery interval of 80-120% (or alternative interval justified in the plan). Some agencies specify more stringent acceptance criteria. For example, DOD recommends acceptance criteria of matrix spike recovery for certain metals analyses to be 85-115%. On the other hand, for certain difficult dioxin compounds the acceptance criteria are 40-135% (https://www.denix.osd.mil/denix/Public/Library/Compliance/EDQW/QSMver2.PDF).**

**For organic compounds and nutrients, a typical performance standard is that lab measurement bias (systematic error, typically quantified as % recovery of a known or expected value normalized to 100% and expressed as comparison % recovery) should not exceed a % recovery interval of 80-120% for a certified reference material. A matrix spike comparison should typically not exceed a % recovery interval of 70-130% (or alternative interval justified in the plan).**

**See also appendix V-F for additional details.**

**V-G. Field Measurement Systematic Error (Bias):**

**Section V of the detailed study plan should briefly summarize how QC performance standards and estimation methods for field measurement systematic error/bias were selected (for example, to be consistent with a listed State or Federal regulatory programs). A statement in the study plan should explain that measurement quality objective and estimation details on this QC topic may be found in the applicable QC SOP included in each protocol.**

**For aquatic monitoring projects, the following guidance should be helpful in guiding what goes into the QC SOP under the heading of field measurement systematic error (bias):**

**Field measurement systematic error (bias) is present when field measurements are consistently too high or too low.**

## Typical Park Service Qualitative Data Quality Objectives:

**The QC performance standards used should be consistent with overall data quality needs based on the data uses. The Measurement Quality performance standards should be specified in the plan, and QC performance results should be specified two ways (as raw data and as summary statistics) in metadata.**

**Unless otherwise justified, performance standards and frequency of QC samples for control of field measurement systematic error (bias) should be at least as stringent as those suggested for EPA estuarine and marine EMAP parameters (http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf).**

**Field measurement systematic error (bias) of biological parameters or biological metrics is sometimes difficult to assess. Nevertheless, one must control systematic error (bias), often by trying multiple methods to see which methods are biased in various directions and then setting logical systematic error (bias) measurement performance standards. For example, a large mesh size may bias the results by excluding smaller organisms, or electro-fishing may collect only certain size classes of fish. For more details, see PBMS discussion of Rapid Bioassessment Precision in Barbour, M.T., J. Gerritsen, B.D. Snyder, and J.B. Stribling. 1999. Rapid**

**Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, Second Edition. EPA 841-B-99-002. U.S. Environmental Protection Agency; Office of Water; Washington, D.C. (http://www.epa.gov/owow/monitoring/rbp/ch04main.html#Section 4.4).**

**Some default quantitative measurement quality objectives are listed below, however, other regulatory or general status and trends monitoring default QC performance standards for field measurement systematic error (bias) may be used when appropriate to ensure data comparability.**

# Typical NPS QUANTITATIVE Measurement Quality Objectives (QC Performance Standards) for Field Measurement Systematic Error (Bias):

**Field measurement systematic error (bias), typically quantified as % recovery of a known-value (normalized to 100% and expressed as comparison % recovery) for a certified reference material, matrix spike sample, or other "known-value" sample, should not exceed a % recovery interval of 80-120%.**

**When possible, systematic error MQOs for field measurement parameters should not be more lenient than Marine EMAP MQOs. This is particularly relevant to marine or estuarine sampling, since all coastal States have standardized on EPA EMAP-Marine protocols; therefore, the group suggests that the EMAP National Coastal Assessment Field Operations Manual protocols (For details, see C. Roman, R. Irwin, R. Curry, M. Kolipinski, J. Portnoy, L. Cameron. 2003. White-Paper Report of the Park Service Vital Signs Workgroup for Monitoring Marine and Estuarine Environments. Workgroup Convened April 3-4, 2002, North Atlantic Coast CESU at the University of Rhode Island, Narragansett, RI. (http://science.nature.nps.gov/im/monitor/COREparamMarine.doc). For example, the Marine EMAP systematic error (bias) QC standards (measurement quality objectives--MQOs) given for field probes, including water column data loggers, include the following expressed as an absolute difference compared to the expected reference material value (EMAP Table A7-1 at http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf):**

**Dissolved oxygen ± 0.5 mg/L**
**Salinity ± 1.0 ppt**
**pH ± 0.3 units**
**Temperature ± 1.0_ C**
**Depth ± 0.5 m (~ 2 ft)**

> **Note from Roy Irwin: if calibration for pH is done with two standards as recommended by EPA, a third standard (NIST or other high quality standard) could be taken to the field for a final check of real-world systematic error/bias.**

**The plan should specify the frequency or number of QC samples.**

**For parameters or estimates to be made in the field, at least one QC systematic error/bias sample should be analyzed in the field prior to analyses of other field samples at that site. This is consistent with EPA guidance to take at least one QC sample in the field prior to field measurements (EPA, 2000, Guidance for assessing chemical contaminant data for use in fish advisories, Volume 1: Fish Sampling and Analysis - Third Edition. EPA 823-B-00-007, Table 8-7, available on the Internet at http://www.epa.gov/ost/fishadvice/volume1/v1ch8.pdf).**

**This recommendation is also consistent with the commonly suggested advice for water column and other samples that QC samples be done for each sampling set (batch), or at a rate of not less than 5%, whichever is more frequent (American Public Health Association, American Water Works Association, and Water Environment Federation. 1998. The 20th Ed. Of Standard Methods for the examination of water and wastewater, 20th Ed. American Public Health Association, Washington, D.C.). A default SOP that has been used by many, including EPA's marine EMAP program, is to limit batch sizes to 20 samples, and to have at least one set of QC samples per each 20 samples (5%). However, when measuring in the field, one is changing batches or conditions each day and/or whenever moving to a new location, so at least one QC sample should be taken at each site and/or each day, even if 20 samples are not taken at that site.**

> **More detail: One often first makes a couple of duplicate measurements to check repeatability precision at a site until two or more consecutive measurements agree or are at least very close. The results of the final two measurements (of exactly the same sample) could be recorded as precision repeatability duplicate measures. This would typically be done before the step being discussed in this section, the step of measuring a known standard to check for systematic error. If one simply moves down the river in short order and measures another site, recalibration and/or another systematic error QC sample may not be necessary until one has reached the 20 sample limit. However, if the duplicate measures do not agree, or if one has put the electronic meters in a pickup and bounced down a rough road to get to another site where chemical and physical (temperature, elevation, pressure, etc.) characteristics may be different, one has essentially started an new sample batch, and new field calibration and new QC samples must be taken. If each sample is in a very different place with different conditions, one may need a QC sample set for each site each day.**

**STORET Notes: QC sample results related to systematic error (bias) and accuracy are assigned to "trips" (individual sampling events). As explained in Part E of this guidance (see Figure 3 and related text), "a trip occurs, for example, when a data collector leaves the office and collects samples and/or makes measurements/observations at the six stations included in the park's monitoring network and then returns to the office. It is included as a way to attach QC data" (Trip Spike, etc.) to all the samples collected during the trip. QC results related to any of these types of samples should be presented, not just summary statistics. Since new STORET relies mainly on a plus or minus field and a confidence interval field in the CHEMICAL DATA RESULT ENTRY BOX, and since not all explanatory information on types of spikes and to the many other possible QC samples related to field measurement systematic error (bias) are listed as standard STORET choices, it is important to list all other important metadata related to systematic error (bias) and/or overall accuracy in the Comments Box that is part of the data result entry boxes.**

**In new STORET, the window for QC Sample (QC6) is the area where one records the results from QC samples including recovery of certified reference materials or spikes (as well as metadata about transportation and storage methods.**

**See appendix V-G for details.**

### V-H. Blank Control Systematic Error (Bias) in Lab Chemical Measurements:

**If applicable (such as when chemicals are being measured in the lab), the detailed monitoring plan should summarize the basics about how blank control will be handled and reported and whose MQOs (if any) are being adopted to help assure data comparability. A statement in the study plan should explain that measurement quality objective and estimation details on this QC topic may be found in the applicable QC SOP included in each protocol.**

**For aquatic monitoring projects, the following guidance should be helpful in guiding what goes into the QC SOP under the heading of lab measurement precision.**

**For parameters that will be detected at some concentration even in pristine environments (such as pH, dissolved oxygen, salinity or specific conductance calculated from conductivity, temperature) and other parameters typically measured with field probes such as turbidity, etc., blank control is not logical, so instead of blank control samples, when possible investigators typically measure certified reference material standards with values close to pristine or locally occurring levels.**

**Until recently blank control was also thought not be logical for chlorophyll a, but recent information suggests that how blanks are handled can make a difference, and that in oceanography, fluorometric determinations of chlorophyll are flawed due to the use of clean water rather than filtered seawater as the method blank (J. J. Cullen and R. F. Davis. 2003. The Blank Can Make A Big Difference In**

**Oceanographic Measurements. Limnology and Oceanography Bulletin 12(2) Internet available at http://www.aslo2.org/bulletin/03_v12_i2.pdf).**

**However, many toxic chemicals are ordinarily not detected in pristine environments in the absence of man's influence. For such contaminants, QC blank control samples are typically used to see if the samples are being accidentally contaminated during collection, during transport, in the lab, or at any other time before the final result is measured.**

# Typical Park Service Qualitative Data Quality Objectives:

**If one finds contamination in "blanks" (samples that should not have detectable quantities of the parameter, being monitoring), one suspects that the samples are being contaminated and makes corrections (or interprets data) accordingly.**

**Field rinsate samples are the minimum recommended when blank control is indicated, but additional types of blank should be used to find the source of contamination if the rinsate blanks show positive detections. Specific types of blank samples are sometimes used to further pinpoint where problems are occurring. For example lab blanks performed on a clean machine can sometimes suggest a calibration problem. However, the more generic field rinsate blanks should typically be performed on at least one sample per sampling trip and/or lab batch (or perhaps lab run, whichever is most applicable). Then if problems are found in those samples, more specific types of blanks could be required to find the source of the contamination.**

**As a default choice, field (trip) blanks used routinely in NPS field studies also be "equipment blanks" in that they are used in the field to rinse collecting equipment. These blanks should also accompany all samples throughout the field trip and be processed, handled, and shipped to the lab like other samples.**

**If the chemical being analyzed is detected in the blank samples, the other data for that sampling trip should either be not reported (rejected) or should be reported with a flag for blank contamination (such as the "V" code consistent with USGS NWIS database code for blank contamination present). Both new STORET and USGS NWIS databases now allow more metadata description space where one can (and should) explain in detail what data flags or codes mean for a given sample. When there is contamination present in blanks, if one is contemplating bounding uncertainty in an environmental concentration or in a measurement process, consideration would have to be given in using net systematic error (bias) in uncertainty calculations (see uncertainty section IV-F.2). For definition of net systematic error (net bias), see definitions section at the end of the appendices.**

**Regulatory or general status and trends monitoring default QC performance standards for blank control other than those listed below may be used when appropriate to ensure data comparability. For example, in CERCLA studies, the plan may specify CERCLA defaults for blank control, and when it is appropriate to attain comparability with USGS NAWQA data, one can use the blank control methods at http://water.wr.usgs.gov/pnsp/pest.rep/sw-t.html#QA). In all cases, the QC performance standards used should be specified in the plan, and QC results should be specified two ways (as raw data and as summary statistics) in metadata. The type of blank should be specified in the STORET QC Sample Data Entry Window.**

> **STORET Note: QC sample results related to systematic error (bias) and accuracy are assigned to "trips" (individual sampling events). As explained in Part E of this guidance, "a trip occurs, for example, when a data collector leaves the office and collects samples and/or makes measurements/observations at the six stations included in the park's monitoring network and then returns to the office. It is included as a way to attach QC data (Trip Blank, Reagent/Method Blank, Equipment Blank, Pre-preservative Blank, Post-preservative Blank in STORET's QC7 QC Sample Data Entry Window) to all the samples collected during the trip. QC results related to any of these types of samples should be presented, not just summary statistics. Since new STORET relies mainly on a plus or minus field and a confidence interval field in the CHEMICAL DATA RESULT ENTRY BOX, and since not all explanatory information on possible types of blanks or other possible field measurements of systematic error (bias) (particularly for biological samples) are listed as standard STORET choices, it is important to list all other important metadata related to systematic error (bias) and/or overall accuracy in the Comments Box that is part of the data result entry box.**

**Studies in which fish tissues are analyzed for contaminant concentrations should use blank quality control least as rigorous as that suggested by EPA (EPA, 2000, Guidance for assessing chemical contaminant data for use in fish advisories, Volume 1: Fish Sampling and Analysis - Third Edition. EPA 823-B-00-007 at http://www.epa.gov/ost/fishadvice/volume1/v1ch8.pdf).**

**It should be kept in mind that blank control systematic error (bias) can be more important (in relation to general data quality objectives) in some cases than in others. For example, if the objectives of the study include characterization of reference conditions that imply accurate low-level measurements of selected constituents are essential, but field blank data indicate systematic low-level systematic error (bias) (sample contamination) occurred during data collection, then obviously describing reference conditions in relation to these data has been compromised. On the other hand, if the objective of the study is to determine whether the targeted constituents had concentrations above some critical value that lies an order of magnitude or larger above values typical of the low-level systematic**

**error (bias) determined by blanks, the effect of the low level contamination on data interpretation objectives my turn out to be comparatively trivial (Michael T. Koterba, USGS, Personal Communication 2002). For more information on how to factor in the effects of both systematic error (bias) and precision on decisions, see DQOs on estimating uncertainty in the measurement process in general and on measurement uncertainty-adjusted confidence intervals) in Step IV-F.2.**

## Typical NPS QUANTITATIVE Measurement Quality Objective (QC Performance Standard) for Blank-Control Measurement Systematic Error (Bias):

**The plan should specify the frequency or number of QC samples for blank control.**

**In the case of marine or estuarine sampling, unless otherwise justified, Performance standards and frequency of QC samples should be at least as stringent as those suggested for EPA estuarine and marine EMAP parameters (for example, reagent (method) blank at the start and completion of a run, limit batch sizes to 20 samples, and to have at least one set of QC samples per each 20 samples (5%). (http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf).**

**Toxic chemical samples in fish tissues or other solids such as soil or sediments should use quality control and frequency of QC method blanks least as rigorous as that suggested by EPA or others generating data to be compared. Unless otherwise justified, a frequency of one method/reagent blank sample per 20 sample batch (5%) is suggested (EPA, 2000, Guidance for assessing chemical contaminant data for use in fish advisories, Volume 1: Fish Sampling and Analysis - Third Edition. EPA 823-B-00-007, Table 8-7, available on the Internet at http://www.epa.gov/ost/fishadvice/volume1/v1ch8.pdf).**

**In other types samples for which blanks are appropriate, QC method blank samples should be done for each sampling set (batch), or at a rate of not less than 5%, whichever is more frequent (American Public Health Association, American Water Works Association, and Water Environment Federation. 1998. The 20th Ed. Of Standard Methods for the examination of water and wastewater, 20th Ed. American Public Health Association, Washington, D.C.).**

**Alternatively, for freshwater samples, in cases where data is to be compared with USGS data, the number of blank samples to be analyzed in the lab compared to total samples should be those suggested by USGS. For details see Table 1 at http://water.usgs.gov/nawqa/protocols/OFR97-223/ofr97-223.pdf. USGS QC methods and terminology are different enough that if USGS data comparability is critical, it might be easier to arrange for USGS to perform all the sampling, lab analyses, and QA/QC.**

**For chemicals that should not be detected in the absence of anthropogenic input, field (field + trip + equipment rinsate) blanks (samples that presumably should remain uncontaminated) should contain no detections of the chemical analyzed above qualitative method detection limits (MDLs). Other acceptance criteria can sometimes be justified. For example, the Department of Defense typically gives blank acceptance criteria as no analytes detected, with none being less than or equal to one half the "reporting limit" (the laboratories lower quantitation limit), (https://www.denix.osd.mil/denix/Public/Library/Compliance/EDQW/QSMver2.PDF).**

**If results are to be corrected for blank contamination, the correction method should be specified.**

> **For example, in one sediment-sampling program, laboratory contamination was corrected by subtracting the 95th-percentile concentration of laboratory blank samples analyzed from the concentration detected in environmental samples (T.J. Lopes and E.T. Furlong. 2000. Occurrence And Potential Adverse Effects Of Semivolatile Organic Compounds In Streambed Sediment, United States, 1992–1995, Environmental Toxicology and Chemistry: Vol. 20, No. 4, pp. 727–737).**

**Note from Roy Irwin: Using the 95th percentile might be fine if one were looking at systematic error (bias) estimates over time, USGS style, say looking at systematic error (bias) in mercury values from 1997 to 2000, as one example. However, for a short term project, or for a situation where one has only data from one lab batch or one run, the best estimate one would have for that single run or lab batch might typically be just one estimate of blank control systematic error (bias) (say plus 2%) and just one estimate of certified reference material (or matrix spike) recovery, of say minus 20% (80% recovery). So for that one run or lab batch, the best estimate one would have of "NET SYSTEMATIC ERROR (BIAS)" would be +2% -20% = minus 18%. For a definition of net systematic error (bias), see definitions section at the end of the apendices.**

**One approach would be to adjust all data in that particular run by –18%. Some data users would adjust data that way. For contrast, USGS tends to not adjust data or to throw out data, but to instead look at frequencies of problems over time, then try to correct for problems accordingly. This is a different approach to systematic error (bias) control. Another approach would be to adjust the data from each lab batch or run according to the best estimates of total measurement uncertainty for that run or batch, then portray long term data as "measurement error-adjusted data."**

**For more information on blank control systematic error (bias) in general, see appendix V-H.**

## Introduction to Accuracy:

**We agree with NIST that "because accuracy is a qualitative concept, one should not use it quantitatively, that is, associate numbers with it; numbers should be associated with measures of uncertainty instead" (http://physics.nist.gov/Document/tn1297.pdf).**

**Likewise, recent EPA guidance suggested using caution in using the word accuracy, a term that has been so misunderstood in the past and so often confused with bias, clarifying that "Accuracy includes a combination of random error (precision) and systematic error (bias) components that are due to sampling and analytical operations"; the EPA therefore recommended using the terms *"precision"* and *"systematic error (bias)*", rather than "accuracy," to convey the information usually associated with accuracy"…" Determination of accuracy always includes the effects of variability (precision); therefore, accuracy is used as a combination of systematic error (bias) and precision (EPA. 1998. EPA Guidance For Quality Assurance Project Plans, EPA QA/G-5, EPA ORD Publication No. EPA/600/R-98/018, http://www.epa.gov/quality/qs-docs/g5-final.pdf).**

**Uncertainty in measurement accuracy can typically be minimized if measurement sensitivity/detection limits are adequate for the decision at hand and if typical sources of measurement error (lack of precision and presence of measurement bias/systematic error) are adequately controlled.**

**Many past documents have confused accuracy and systematic error (bias). However, it is clear that the concept of accuracy includes the contributions from both precision and systematic error (bias). The more recent EPA QA/QC guidance documents have included illustrations that show to be accurate data must be both precise and unbiased. Using the analogy of archery, to be accurate, one must have one's arrows land close together (reflecting a high degree of precision) and, on average, at the spot where they are aimed (neither consistently high nor low, reflecting a low degree systematic error (bias). That is, the arrows must all land near the bull's-eye (EPA. 1998. EPA Guidance For Quality Assurance Project Plans, EPA QA/G-5, EPA ORD Publication No. EPA/600/R-98/018, http://www.epa.gov/quality/qs-docs/g5-final.pdf, see also illustration in Guidance for Volunteers on page 26 of http://www.epa.gov/volunteer/qapp/vol_qapp.pdf).**

**Nevertheless, some recent EPA documents prolong confusion by treating systematic error (bias) and accuracy as synonyms (EPA. 2001. Environmental Monitoring and Assessment Program (EMAP): National Coastal Assessment Quality Assurance Project Plan 2001-2004. United States Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Gulf Ecology Division, Gulf Breeze, FL. EPA/620/R-01/002. (http://www.epa.gov/emap/nca/html/docs/c2k_qapp.pdf).**

**Different agencies handle the concept of accuracy differently. For example, the USGS NAWQA QC design summary document doesn't even mention the word accuracy, but recognizes the concept by saying "To interpret water-quality data properly, information is needed to estimate the systematic error (bias) and variability that result from sample collection" (http://water.usgs.gov/nawqa/protocols/OFR97-223/ofr97-223.pdf).**

**In another USGS paper, the word variability is used instead of precision and it is stated, "Most authorities agree that separate narrative statements of random error (variability) and systematic error are required" to express uncertainty in a measurement process, and that "a probability interpretation (such as a level of confidence) is desirable "(J.D. Martin, 2002. Variability of Pesticide Detections and Concentrations in Field Replicate Water Samples Collected for the National Water-Quality Assessment Program, 1992-97, Water Resources Investigation Report 01-4178. NAWQA, Indianapolis, IN, 84 pages, http://water.usgs.gov/pubs/immediate_release.html). One may recognize the link between uncertainty and accuracy, in that both concepts require one to consider both variability (precision) and systematic error (bias). For more details on how to estimate confidence (or its inverse, uncertainty) quantitatively after considering factors related to both precision and systematic error (bias), see separate step IV-F.2 on bounding error or uncertainty.**

### V-I. Uncertainty in Accuracy

**The detailed study plan should summarize the basics about the degree to which uncertainty in overall measurement accuracy will be controlled and reported. A statement in the study plan should explain that measurement quality objective and estimation details on this QC topic may be found in the applicable QC SOP included in each protocol.**

**For aquatic monitoring projects, the following guidance should be helpful in guiding what goes into the QC SOP under the heading of uncertainty in accuracy.**

## Typical Park Service Qualitative Data Quality Objectives For Uncertainty in Accuracy:

**The uncertainty associated with sample observations or measurement should be as low as possible. In typical scenarios, this translates to the concept that systematic error (bias) should be minimized and measurement error variability (expressed as lack of measurement precision) should be minimized. This is a qualitative data quality objective.**

**Measurement uncertainty about single data points should be estimated as an NIST-defined expanded uncertainty. Uncertainty in means and other summary statistics should be estimated as a measurement uncertainty confidence interval (see section IV-F.2).**

**There is no generic quantitative data quality objective performance standard for accuracy as a whole. However, most of the time, if data quality objectives and performance standards for precision and systematic error (bias) are met separately, uncertainty in measurement accuracy will be acceptable.**

**Therefore, there are no required data quality objectives for (uncertainty in) accuracy per say, but rather DQOs and performance specifications for the main individual contributors to uncertainty: data quality indicators like sensitivity, precision, reference material and/or spiked sample systematic error (bias), and blank control systematic error (bias) (see separate sections herein on these topics).**

**However, if planners wish to develop Data Quality objectives for uncertainty as a stand alone, it is suggested that they use the Root Accuracy Square Error (RSS) uncertainty approach as detailed as detailed in appendix IV-F.2).**

**STORET Note: QC sample results related to systematic error (bias) and accuracy are assigned to "trips" (individual sampling events), which would often coincide with a lab batch or run. As explained in Part E of this guidance, "a trip occurs, for example, when a data collector leaves the office and collects samples and/or makes measurements/observations at the six stations included in the park's monitoring network and then returns to the office. It is included as a way to attach QC data" (Trip Blank, Reagent/Method Blank, Equipment Blank, Pre-preservative Blank, Post-preservative Blank, Reference Sample, Trip Spike) to all the samples collected during the trip. QC results related to any of these types of samples should be presented, not just summary statistics. Systematic error (bias) calculations vs. certified reference samples (CRMs, also called a laboratory control standard by some labs) are not well provided for in version 1.2 of STORET, an issue that will hopefully be corrected in the newer version coming out in the summer of 2002. At the moment, the main place to address systematic error (bias) is in the QC adjustment factors area, see window APL2, where EPA seems to be (wrongly) assuming that most labs report systematic error (bias)-adjusted results when recoveries are not 100%. Until the newer STORET comes out, it is especially important to list all other important metadata related to systematic error (bias) and/or overall accuracy in the Comments Box that is part of the chemical data result entry box. In the CHEMICAL DATA RESULT ENTRY BOX, there is a plus or minus field given for precision. Next to that is a confidence interval field for precision. A confidence interval is not mandatory and should not be calculated if sample size is less than 10. If the confidence interval is a systematic error (bias)-adjusted, or better yet a confidence interval adjusted for total measurement uncertainty, that fact should be explained in the comments box, especially since standard STORET choices give confidence intervals for precision results but not for accuracy (a combination of systematic error (bias) and precision).**

# Controlling Uncertainty in Overall Accuracy with Rounding Rules:

**Rounding Rules --How Many Significant Figures Should be Reported?**

**Those reporting or archiving data (such as temperature, pH, and conductivity) that will commonly later be used in the calculation of another value, should be careful not to round so aggressively that they inadvertently contribute to cumulative rounding errors in later calculations. A typical example of historically taught rules of thumb is provided on a John Hopkins engineering school website (http://www.apl.jhu.edu/Classes/Notes/Telford/SignificantDigits.pdf:**

1. **In your calculations, use one or two extra digits in intermediate calculations to avoid round-off errors and then round off the final result appropriately.**

2. **What is appropriate rounding off of the final result?**

   **When performing additions and subtractions only, the final result may have no more significant digits after the decimal point than the number with the fewest significant digits after the decimal point.**

   **When performing multiplication and divisions only, the final result may have no more significant digits than the number with the fewest significant digits (ignore the decimal point).**

**Although rounding rules of thumb such as those quoted above have value, there is no single answer on the best way to determine the number of significant figures to use for all applications.**

To some degree, many of the widely taught simplistic rounding rules reflected older (slide-rule era) thinking on how to at least crudely factor uncertainty into a final result. A community college website provides a plain language explanation of how rounding considerations relate to measurement precision and to uncertainty in final results:

> “A measurement reported as 45.67 mL indicates that the piece of equipment being used could be measured precisely to 0.1 mL and a reasonable guess can be made about the hundredths place. So, this measurement is taken as being somewhere between 45.66 and 45.68 mL. There is some uncertainty in that last decimal place, but since a reasonable estimate can be made for its value, the 7 is still significant. Using the same piece of equipment, it would be dishonest to report a measurement of 32.446 mL because the equipment can only measure out to ± 0.01 mL. Any decimal places beyond the hundredths place would be pure guesses, and not at all significant for the value of this measurement” (http://www2.aacc.cc.md.us/sciljtracey/CHE111/c111helpsigfig.htm)

**Historically, cumulative and other rounding errors in final results have been common in general environmental work. They can also arise when different individuals involved with a project do not round consistently. Workers seldom have been able to justify going beyond slide rule accuracy (three significant digits) for final "results." In fact, it is not unusual to conclude that not more than two significant figures are truly justified (Owen Hoffman, SENES Oak Ridge Inc., Personal Communication, 2000). For hydraulic conductivity and "storage coefficient" groundwater parameters, it has often hard to justify more than one significant figure (Bill Van Liew, NPS, Personal Communication, 2002). Some chemical analytical labs have used a "report no more than three significant figures" as a rule of thumb applicable in at least some scenarios (Roy-Keith Smith, 1997. Laboratory Analyst Training un the 1990's And Beyond. WTQA '97 - 13th Annual Waste Testing & Quality Assurance Symposium, http://www.clu-in.org/download/char/dataquality/rsmith.pdf).**

**Why have investigators in the past so often worried about rounding final results only to justified numbers of significant figures? Typically, the answer is that unless measurement uncertainty is calculated, sensitivity and overall uncertainty in accuracy (after considering both precision and bias) of environmental measurement devices do not justify reporting larger numbers of digits.**

**ASTM acknowledges one example of this in its reporting suggestions for turbidity, where no more than two significant figures are recommended for results up to 1000 NTU units. Above 1000 NTU, up to 3 significant figures can sometimes be justified (ASTM. 1988. Standard Test Method for Turbidity in Water, D 1889-88A).**

**Model inputs given to too many significant figures (often beyond three) can give a misleading impression of precision. In the presence of uncertainty, it is better to define all uncertain inputs as subjective probability distributions (no need to worry about rounding) and propagate this information through a model using Monte Carlo simulation to produce a probability distribution for the result. If one uses the Monte Carlo error propagation technique, the computer uses a very large number of digits for all values sampled within the specified limits of the distribution given to represent the state of knowledge of the true but unknown value. I don't think rounding errors are an issue when full accounting of uncertainty is considered. (Owen Hoffman, SENES Oak Ridge Inc., Personal Communication, 2000). The 20th edition of Standard Methods (American Public Health Association, American Water Works Association, and Water Environment Federation. 1998. The 20th Ed. Of Standard Methods for the examination of water and wastewater, 20th Ed. American Public Health Association, Washington, D.C.) is the first edition to recognize Monte Carlo and other advanced methods to bound uncertainty.**

**With the advent of modern ways to express uncertainty and computers that will carry large numbers of "guard" digits to protect against cumulative rounding errors in subsequent calculations, we now have better ways to express uncertainty in final results. In the example given above, which takes into account only**

**measurement precision, the value was rounded to 45.67. A more modern and more quantitative way to express uncertainty would be to use NIST measurement uncertainty calculations that take into account not only precision, but also other known contributors to uncertainty, such as measurement systematic error/ bias (N. Taylor and C. E. Kuyatt. 1994. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results NIST Publication TN 1297 (http://physics.nist.gov/Document/tn1297.pdf). If one were using this method to express uncertainty in a final result rather than simply rounding a value to 45.67, one would instead calculate and express measurement uncertainty in the final result as an NIST measurement uncertainty plus or minus value to accompany the result. For example, if measurement precision was very good and the total measurement uncertainty was very low, such a result might be expressed as something like the following: 45.676 plus or minus the measurement uncertainty, perhaps something like 0.003. If the measurement uncertainty for a single measurement was very high (not so unusual in some environmental observations or measurements), the single point might be reported as something like 50 plus or minus 30. Uncertainty in a mean can be expressed in a properly framed confidence interval about the mean, lengthened by the amount measurement uncertainty about each data point.(for more details, see Park Service monitoring guidance at http://science.nature.nps.gov/im/monitor/protocols/wq)artB.doc.**

**However, data reporters do not always know how the data they produce will be used in subsequent calculations. Data users perhaps need to pay even more attention to uncertainty and to rounding than data reporters. If one is trying to determine if there are trends or if one is trying to determine if one set of numbers is statistically different than another, one should typically use measurement uncertainty adjusted confidence intervals as discussed above.**

**A much more crude way and old fashioned way to account for uncertainty in a single data point, but probably better than not accounting for uncertainty at all, would be to use a rounding rule such as the one below, for final results based on factors such as measurement precision as reproducibility.:**

> **First determine the number of significant figures one could round the precision reproducibility values (repeat observations on the same substance by different persons or using different instruments) so that the rounded value reported would not change between observations. Next, add one to that number of significant figures, so that there is some uncertainty only in that last significant figure or decimal place but so that the last significant figure is a reasonable estimate and not a pure guess.**
>
> **An example of this kind of rounding is provided as follows:**
>
> > **Those who resist the best practice of calculating measurement uncertainty could perhaps use the probe makes measurement precision specifications to makes these estimates. However, the actual**

**precision one can achieve in the field or lab is often not as good as that given by the probe manufacturer. For example, let's take a look at some actual precision numbers for Specific Conductance. Using the rounding rule above (certainty plus one significant figure), I have put the number of significant figures (SF) beyond which we have uncertainty in parenthesis behind each of the sets of numbers:**

| **Trials:** | **Precision Values** | | | **(round to SF)** | **Rounded Values** |
|---|---|---|---|---|---|
| **Soln 1** | **142** | **131** | **139** | **(2)** | **140, 130, 14** |
| **Soln 2** | **1866** | **1850** | **1844** | **(2)** | **1900,1800, 1800** |
| **Soln 3** | **1872** | **1850** | **1854** | **(2)** | **1900, 1800, 1900** |
| **Soln 4** | **1865** | **1860** | **1869** | **(2)** | **1860, 1860, 1870** |
| **Soln 5** | **1235** | **1227** | **1254** | **(2)** | **1200, 1200, 1300** |
| **Soln 6** | **222** | **217** | **223** | **(3)** | **222, 217, 223** |
| **Soln 7** | **4282** | **4430** | **4420** | **(2)** | **4300, 4400, 4400** |
| **Soln 8** | **259** | **253** | **258** | **(2)** | **260, 250, 260** |
| **Soln 9** | **1970** | **1970** | **1943** | **(2)** | **2000, 2000, 1900** |
| **Soln 10** | **12,392** | **12,020** | **12,629** | **(2)** | **1200,1200,1300** |

**This concept of reporting only one significant figure beyond what is known definitely in final results, the number of significant figures to be used in multiplication or division, and how to round when a standard deviation is known, are all explained in more detail in Section 1050 B "Significant Figures" of Standard Methods (American Public Health Association, American Water Works Association, and Water Environment Federation. 1998. Standard methods for the examination of water and wastewater, 20th Edition. American Public Health Association, Washington, D.C.), a reference that many chemical labs have on hand.**

**One relatively simple approach to avoiding bias related to cumulative rounding errors is to use rounding calculators that round up half the time and down half the time when the last digit before rounding is five, based on whether the previous digit is even or odd (http://ostermiller.org/calc/sigfig.html).**

**It might be especially tempting to use a rounding rules based on precision only, such as the one above, to account for uncertainty in parameters that don't lend themselves to easy estimates of systematic error (bias), such as:**

1. **PAR**
2. **Bacterial Counts**
3. **Taxonomic Identification of Very Small Invertebrates or Other Difficult Taxa**
4. **Judgment Habitat Observations (Percent Embededness of Oysters)**
5. **Spike Recoveries of Chemicals in Difficult Matrices**
6. **Dissolved Oxygen Measurements**

**In these types of parameters, the expected or correct answer is not always easy to identify, so systematic error (bias) is more difficult to estimate and it might be more tempting to confine uncertainty to precision (repeatability) aspects or even rounding rules only. However, even for these types of parameters, there is often some way to at least roughly estimate systematic error (bias). For example, sometimes the observation or estimate of an expert is considered "right" or "expected" result and the difference between that observation and those of rookies or trainees is considered systematic error (bias). In cases where one is not sure even an expert is "right", another approach would be to take the maximum difference (delta) between observations as systematic error (bias). In this case, the systematic error (bias) estimate would be the same as maximum difference in reproducibility (something, like the person doing it or the instrument changes) precision. That would then be the conservative (worst case) estimate of bias and the variance of that value would be added to the variance of the value for precision repeatability (nothing changes) in sum of squares NIST calculations of measurement uncertainty. This approach would perhaps overestimate systematic error (bias) and express it as a plus or minus factor, perhaps not a bad thing when the right answer is not easy to pin down. An approach such as this one would allow one to estimate NIST measurement uncertainty, and in most cases would still be superior to trying to use rounding rules as crude ways to account for uncertainty.**

**For those who wish to study rounding rules issues further, the plain language explanation and examples provided by John S. Denker (http://www.av8n.com/physics/measurement-u.htm#bib-basic) are recommended. These discussions:**

1. **Are well written,**
2. **Explain why there is no single best way to choose the number of significant figures justified,**
3. **Explain that NIST measurement is a better way to document uncertainty in the final result than rounding rules.**
4. **Explain how excess rounding can result in important cumulative rounding errors,**
5. **Explain how excess rounding can hinder picking signals out of noise, and**
6. **Provide spreadsheet examples relevant to signal to noise ratios. Signal to noise ratios are very relevant to detection limits, deciding when a stressor impact is causing a change considered beyond normal, and other important issues relevant to environmental monitoring.**

**In summary, there are hazards from excess rounding of intermediate values, and modern computers negate some of the reasons we previously had for doing so. The best way to express uncertainty in final results is by providing an NIST suggested calculation of measurement uncertainty. For those who are unwilling to take the time to do this, the alternative of rounding a final result single data point based on limits of precision repeatability, is perhaps better than nothing, and may be tempting for those parameters for which estimates of systematic error are difficult,**

**but this option is almost always far less complete or quantitative than estimating total measurement uncertainty.**

**Finally, none of these considerations negates the need for data users to consider common sense rounding rules related to multiplication/division and adding and subtracting, such as those mentioned above.**

**Data Quality Objectives:**

**Any rounding rules to be used shall be detailed in the plan. Rounding errors and cumulative rounding errors shall be minimized.**

**Using rounding to add an aspect of uncertainty in final results shall not be done take the place of calculating measurement uncertainty, and the rationales used for any rounding rules shall be justified in the plan.**

## Uncertainty in Software Accuracy:

**Disclaimer: No government endorsement is implied for software examples mentioned. No representation is made that all worthy software alternatives are mentioned:**

**Several recent publications have discussed potential issues with accuracy of statistical calculations in MS Excel, including Cox (2000), Cryer (2001), Knusel (1998), McCullough (1998) McCullough (1999), and McCullough and Wilson (1999). Considering the findings of these publications and the advice of some statistical experts, what should scientists in the Park Service know? Are calculations in MS Excel, the Park Service "standard spreadsheet", reliable enough for scientific credibility purposes? Should one go with the strong condemnations of Cryer (2001) "Friends Don't Let Friends Use Excel for Statistics!" or the more subdued advice of McBride (see details below) "MS Excel "is fine for BASIC calculations, except percentiles and odd data sets"?**

**After reading the papers cited above and discussing the issues with independent statistical experts, I have decided that the answer depends on the situation and therefore it is difficult to provide blanket "one size fits all" guidance. Instead, I will attempt to alert Park Service scientists to some of the issues and summarize at least some of "the basics" that Park Service Scientists should know.**

**First, we need to at least briefly consider how we might logically define "accurate enough" in the context of software calculations. Uncertainty in accuracy of software is just one of many known sources of uncertainty. Another source is measurement uncertainty. For environmental measurements, measurement uncertainty (factoring in both precision and systematic error/bias) is seldom lower than a plus or minus 3% and is often much higher for parameters like pesticide concentrations in sediments or tissues or for observations such as "percent embededness of cobbles."**

**NIST publishes ISO-compatible sum of squares equations for combining uncertainty from many sources (N. Taylor and C. E. Kuyatt. 1994.Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results NIST Publication TN 1297 (http://physics.nist.gov/Document/tn1297.pdf). In these equations, if the (standard deviation) contributor to uncertainty from one source is five times lower than another contributor, it is considered trivial and not even considered in the overall uncertainty equation.**

**In one of the MS Excel-critical publications (McCullough Wilson, 1999), a standard deviation calculated by MS Excel was 0.0790105482336451. This was compared to a "correct" standard deviation of 0.0790105478190518 calculated by a benchmark standard. Most environmental specialists would round these numbers to the same value (say 0.079). Even if one chooses not to round the numbers prior to their use in equations, the difference between the two doesn't approach one fifth of typical real world environmental measurement uncertainty, and thus it would usually be considered trivial and not added to overall uncertainty equations. Furthermore, measurement uncertainty itself is seldom the source of the greatest amount of overall uncertainty. Model uncertainty (see section IV-F.2), uncertainty in representativeness of the sample, and errors arising from not using software correctly are common. In fact, using software incorrectly may sometimes be more common when using dedicated statistical software than when using the more "user-friendly" and more ubiquitous MS Excel. Other common sources of errors and uncertainty include errors caused by choosing the wrong analysis, errors related to not meeting critical assumptions, and cumulative rounding errors. Furthermore, there are inherent uncertainties related to our imperfect knowledge of biology and physical science, wrong or crude theories, and various sampling errors. Such errors typically vary in magnitude in both time and space. Collectively, these "additional" sources of uncertainty are probably of greater magnitude than software calculation errors, particularly for simple summary statistics.**

**In fact, I have not yet seen any examples where MS gave an answer for a simple statistics such a mean or a sample standard deviation "different enough" to be considered non-trivial in comparison with other sources of uncertainty in environmental variables. Given the fact that all Park Service offices typically already have MS Excel, some would advise that for very simple calculations, like a population standard deviation, a mean, or even a 95% "t distribution" confidence interval, answers in MS may be "accurate enough" for the purposes for the environmental data sets being analyzed and for the statistics commonly used. In these situations, in may difficult to justify having multiple Parks invest in dedicated statistical software.**

**Graham McBride uses and deems MS Excel useful not only for some routine analyses, but also for less-routine multiple one-sided (TOST) tests for inequivalence and some Bayesian statistics that use routine functions. Graham points out that MS Excel is fine for basic calculations such as means, standard deviations, and t values, and that most of the criticisms of Excel are valid only for very odd data, Although Excel does seem to have some problems with percentiles, so does S-Plus. Except for SAS, stats packages typically do not explain that there is no one "right" way to calculate percentiles, let alone tell you which one they use. Users shouldn't use any**

**computer package blindly and users should seek expert statistical advice when needed. The level of explanation in manuals and help files is often poor (Graham McBride, National Institute of Water and Atmospheric Research, New Zealand, Personal Communication, 2002).**

**However, there is probably greater risk from using MS Excel to calculate more complicated statistics such as regression statistics. For complicated statistics, MS Excel calculations may sometimes not be acceptably accurate and the investment in dedicated statistical software would be more easily defended, especially in legally or professionally contentious settings. Keep in mind however, that even dedicated statistical software packages can choke on very difficult data sets.**

**What are difficult data sets? They include those that involve large numbers (usually greater than 6 digits)], and/or a very large sample size (high n), and/or constant leading digits (90000001, 90000002, and 90000003, for example).**

**Odd or difficult data sets are typically rare in environmental work, but this may change as more and more continuous data readout probes are used and the Park Service accumulates long term data sets from the Vital Signs and other long term monitoring programs.**

**What sample sizes are too large? If the individual numbers are high six digits, and the sample size is anything but small, the data set may be starting to "become difficult" even for the relatively simple standard deviation determinations in MS Excel. However, it is very hard for ordinary users to determine case-by-case limitations. It would be nice if the soft makers explained the following for each software/hardware combination: if the sample size is less than "n" and no numbers have more than "z" digits, then large sample/large number-related problems will not arise (Bruce McCullough, FCC, Personal Communication, 2000). They don't,**

**In cases where percentiles or complicated statistics are calculated in MS Excel, it is recommended that for quality assurance purposes, data analyses should be replicated on at least one "dedicated" statistical software program to help insure accuracy.**

**This "other" software should be one that is dedicated to statistical tasks, such as the following (not necessarily complete, no particular endorsement implied, provided as examples) list of typical or widely used examples: SAS, SPSS, SYSTAT, MATHEMATICA, EquivTest, WQStat, MINITAB, STATGRAPHICS, STATA, MAPLE, or S-PLUS.**

**For those who desire to work in the familiar format of MS. Excel, there numerous Excel add-ins, some for beginners (such as add-ins that come with the book Practical Statistics Using Microsoft Excel and Minitab (http://www.amazon.com/exec/obidos/tg/detail/-/0130415219/ref=pd_sim_books_1/102-9188059-8564102?v=glance&s=books). A potential issue with some of these is how accurate some of the add-ins are, and one can seemingly find internet criticisms of certain functions related to many add-ins. However, there are also a vary wide variety other Excel add-in tools which do more complex tasks, such as re-sampling and Monte Carlo Analysis. Some of these add-ins state they have replaced functions in Excel with more robust and accurate functions (for example see StatTools at http://www.palisade.com/html/stattools.asp).**

**A logical first "additional" quality assurance step related to software accuracy when complicated statistics are to be calculated, or when very large or otherwise difficult data sets are to be analyzed, or when legal or rigorous professional challenges are expected, would be to compare results of analyses of NIST published standard data sets with certified NIST correct answers for calculations of the same type as will be performed for the work at hand. Such data sets and certified answers are published by the federal NIST at http://www.itl.nist.gov/div898/strd/.**

**One thing to keep in mind is the more complicated the statistic, the more can go wrong, so we should not be surprised to see more software errors of multivariate or other complicated procedures than in calculations of the sample mean. McCullough's 1999 summary confirms more errors on the relatively complicated (multi-factor ANOVAs and nonlinear) statistical procedures than on the relatively simple univariate procedures (B.D. McCullough. 1999. Assessing the Reliability of statistical software, Part II. The American Statistician 53: 149-159).**

**Cox (2000) likewise suggested that areas where Excel may be unreliable include some relatively complex tasks and/or unusual data sets:**

- **Standard deviations and statistics (eg t-tests) relying on standard deviation calculations where there are large numbers with low variation**
- **Multiple regression with very high collinearity.**
- **Non-linear regression problems.**
- **Distribution tail areas beyond about $10^{-6}$.**
- **Procedures (e.g. bootstrap) that rely on a good random number generator**

**Cryer (2001) offers one of the more strongly worded cautions against using MS Excel (including some newer versions) for statistics, citing not only issues with regression analyses but also with graphing functions and even with the algorithms used to calculate simple summary statistics such as a standard deviation. Cryer points out that Excel has problems with how it treats missing data and ties and (among other things) often displays many more digits than are warranted,**

**However, the problem of displaying too many digits is not limited to Excel. Some programs allow one to specify the number of digits reported, and in any case, it is typically up to users to use logical rounding rules, such as those summarized herein in section V-I.**

**In trials for calculating a sample standard deviation with MS Excel and other dedicated statistical software (Systat), both programs usually tended to return the same standard deviation when rounded to a reasonable number of digits, so it has not yet been demonstrated that there is a major problem with the way Excel calculates sample standard deviations for typical environmental data sets.**

**What about reliability of other software programs? Although potential problems with MS Excel have been well publicized, it is less broadly understood that other software and software/hardware combinations can have problems or limitations too. Even dedicated statistical software programs can choke on very difficult or contrived data sets. Using relatively difficult tests in 1999, flaws were**

**discovered in the then-current versions of widely used statistical packages such as those from SAS, SPSS, S-Plus. SPSS and S-Plus (McCullough, 1999). In a separate comparison involving relatively simple tests, some problems were noted with some of the dedicated statistical packages, as well as generally more serious problems with MS Excel (Landwehr and Tasker, 1999).**

**The picture gets further muddied when considering various multivariate, ordination, and phylogenetic classification programs. For example, taxonomists use such programs to help classify the phylogenetic relationships between species. Countless schemes are used, and some programs are developed by individuals and have very little documentation. Users should be aware that the answers may be suspect and are not always consistent between different programs.**

**For example, in one trial, taxonomists tried several popular multivariate software programs on identical data sets and came up with very different answers. When they tried it again entering data in a different order, they got still yet other answers (Terry Frest, Consultant and malacologist, Personal Communication, 2000).**

**It is difficult to control a measurement process without controlling for systematic error (bias). Bias is difficult to estimate if one cannot identify what the right, or at least the "expected" answer is. This makes quantifying uncertainty in the answers from many multivariate, ordinations, and classification programs very difficult. Typically only type B (expert opinion, non-statistical, and often qualitative) estimates of uncertainty are possible. For additional details on estimating measurement and simple types of model uncertainty, see section IV-F.2 and IV-F.3.**

**References Cited in Section on Uncertainty in Software Accuracy (above):**

**Cox, N. (2000). Use of Excel for Statistical Analysis, Published on Web at** http://www.agresearch.cri.nz/Science/Statistics/exceluse.htm.

**Cryer, J. (2001) Problems With Using Microsoft Excel for Statistics**
**Paper Given at Joint Statistical Meetings in Atlanta, GA, Summary on Web at http://www.stat.uiowa.edu/~jcryer/JSMTalk2001.pdf).**

**Knusel, L., (1998) On the accuracy of statistical distributions in Microsoft Excel 97. Computational Statistics and Data Analysis 26, 375-377 (this plus other relevant contributions by Knusel may be found at http://www.stat.uni-muenchen.de/~knuesel/elv/accuracy.html).**

**J.M. Landwehr and G.D. Tasker (1999). Notes on numerical reliability of several statistical analysis programs. USGS Open File Report 99-95, Reston, VA (http://water.usgs.gov/pubs/of/ofr99-95/ofr.99-95.pdf).**

**McCullough, B.D., (1998) Assessing the reliability of statistical software: Part I. The American Statistician 52, 358-366. (http://www.amstat.org/publications/tas/mccull-1.pdf).**

**McCullough, B.D., (1999) Assessing the reliability of statistical software: Part II. The American Statistician 53, 149-159 (http://www.amstat.org/publications/tas/mccull.pdf).**

**McCullough B.D. and Wilson B., (1999) On the accuracy of statistical procedures in Microsoft Excel 97. Computational Statistics and Data Analysis 31, 27-37 (http://www.seismo.unr.edu/ftp/pub/updates/louie/mccullough.pdf).**

## Field Probe Uncertainty in Accuracy:

**As discussed in more detail in Section V-B.3, monitoring planners should be aware that what field-measurement probe manufacturers mean by "accuracy" tends to vary between manufacturers and tends to be as optimistic as possible for the sake of equaling or better the specifications published by competitors. Both the so-called accuracy and resolution "specifications" reported by manufacturers tend to be ill-defined and too often based on ideal, controlled lab conditions not realistically achievable in field deployments.**

**It would be logical to replace accuracy specifications with NIST/ISO expanded measurement uncertainty (N. Taylor and C. E. Kuyatt. 1994. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results NIST Publication TN 1297 (http://physics.nist.gov/Document/tn1297.pdf) for these specifications within the applicable measurement ranges listed for each device. However, we know of no examples where this has been done. At the present, at least one manufacturer seems to be considering both precision and systematic error (but not in the standard NIST ways), others seem to be considering only precision, and it is not clear exactly how others are developing accuracy specifications.**

**It is best to determine site-specific performance for accuracy by determining both precision and systematic error (bias) IN THE FIELD following final calibration IN THE FIELD. Precision and systematic error (bias) should be measured in the field environment.**

**In any case, QC uncertainty in accuracy objectives (the real-world values from the table above might be used as a pre-study default estimate in the absence of better data) should be detailed in the plan, and actual QC performance should be reported as results metadata.**

**More Detail Related to Field Probes:**

> **Probe users should keep in the mind that the true accuracy (precision plus systematic error (bias) that a user can get IN THE FIELD will almost undoubtedly be different than the manufacturers specifications, which tend to be generated under ideal lab conditions, especially if the manufacturer's calibration instructions are not followed very closely. In general, what probe**

> **makers mean by accuracy can be different than what other probe makers mean or what most chemical laboratories mean. For example, a Hydrolab brochure states that the accuracy of the Quanta meter for turbidity is "plus or minus 5% of reading plus or minus 1 NTU." What does this mean? Even after repeated attempts to explain what it meant, Hydrolab had not made it clear (See appendix section IV-F.2, discussion of Quanta probe for details).**

**If the performance standard stated in the plan is the accuracy standard given by the probe manufacturer (for example for one YSI oxygen meter is +/-2% or +/-0.2 mg/L which ever is greater), checks should be done to see how well the instrument actually performs in the field after field calibration.**

**After very careful calibration according to the manufacturer's recommendations, final estimates of random error (precision) and systematic error (bias) performance typically should be done in the field prior to field measurement. This should be done after allowing the standards and measuring instrument to adjust to field conditions (including temperature). Repeated measurements of a regular calibration standard (or even a homogeneous, well mixed water sample) should be done to get an estimate of precision repeatability, and repeated measurements of an extra (new, third or fourth) NIST approved or NIST traceable standard should be done to get an estimate of the systematic error (bias) performance of the probe following final calibration steps. If a continuous monitoring probe is to be used, in addition to checking for precision and bias performance before deployment, one should also recheck the QC performance of the probe at the end of the deployment period.**

**In real world conditions in the field, it would not be surprising if the performance does not fall into the interval of the probe maker's specification expressed as a plus or minus range for the "accuracy specification" or even for the more rigorous NIST expanded uncertainty. What makes sense depends partly on protocol that is being followed and the manufacturer's suggestions. For example most pH meters recommend multipoint calibration with standards. The USGS recommends 2-3 points be used in calibration, DOE 2, EMAP estuarine program recommends 2,and some lab experts recommend 6 for certain critical applications such as global warming-related measurements of seawater. After one has finished final calibration steps, one should then check instrument performance against an additional (new) NIST traceable standard to get an estimate of systematic error. If the manufacturer of the standard provides it, the uncertainty variance from the standard itself not being perfect could be added to precision variance and systematic error (bias) variance to get a sum of squares. The root of the sum could then be used to get a real world NIST-recommended root sum of squares calculation of "combined uncertainty" (see step IV-F.2). In some cases of unusual (low alkalinity water, weather extremes, etc.) water, field calibrations might not be able to achieve 0.02 for combined uncertainty, and one could then specify and "practical" field performance standard of say plus or minus 0.1 pH, similar to Department of Energy (DOE) protocols. Field in-situ pH measurements are usually rounded to the nearest 0.1 pH unit. However, one often does not need (and is often not justified due to precision**

**and systematic error (bias) limitations in field work) in reporting more than one-tenth pH unit of resolution anyway. It is often difficult to justify rounding most environmental variables to more than 2 or three significant figures. Furthermore, for QA/QC purposes in fieldwork, it is not always so critical if the performance standard systematic error (bias) or combined uncertainty very low, say 0.02 pH units for example. Achieving 0.02 pH units for NIST combined uncertainty may impractical or impossible to in general environmental field monitoring. What is important is not necessarily that uncertainty be very low, but that precision and systematic error (bias) performance standards (and or combined uncertainty) be specified and standardized, and thus both random (precision) and systematic (bias) error be "controlled" and that the QC performance standards picked be and that they are appropriate for the intended uses of the data (sometimes the protocol or regulatory purpose defines minimum performance standards for precision and bias).**

**See appendix V-I for more details. See also separate Part A guidance on calibration of field probes (http://science.nature.nps.gov/im/monitor/protocols/wqPartA.doc).**

## VI. DATA MANAGEMENT

### VI-A. Data Management And Handling

**Monitoring information is not useful if it is not reported, is lost, is not reported in enough detail, or is not available to users.**

**A study of flawed/failed monitoring projects revealed that many problems could have been avoided if sufficient collateral information needed to interpret the data had been reported (L.M. Reid. 2001, The epidemiology of monitoring. Jour. Amer. Water Resources Assn. 37(4): 815-819).**

**Typical NPS qualitative data handling, data management, information reporting, and data archiving objectives:**

**In accordance with general I&M guidance, the plan should include the following items (http://science.nature.nps.gov/im/monitor/monplan.doc):**

- **An overview of the process for entering, editing, storing, archiving, and analyzing data collected by the various components of the monitoring program, including metadata procedures. The full Data Management Plan should be attached as an appendix.**
- **An overview of the database design for the monitoring program.**
- **A description of the various reports and other products of the monitoring effort, including what they will include, who the intended audience is, how often they will be produced and in what format, and who will be responsible for ensuring that data are analyzed and reported in a timely manner.**

> **Helpful document: See also the new guidance entitled "Chapter X: Product Specifications" being developed by the NPS I&M group (see working draft at http://science.nature.nps.gov/im/apps/specs/index.htm).**

**All data results and accompanying metadata should be reported to the Park and should be archived in the new STORET database. For more detail, see Section E of this document (Guidance on Data Reporting and Archiving in STORET) at (http://science.nature.nps.gov/im/monitor/protocols/wqPartE.doc).**

**Metadata details describing "exactly what was done in the field and the lab" should be included in the plan and in STORET metadata fields in order to meet another NPS "good science" data quality objective basic, that measurements of an unchanging substance (for example, a certified reference material should be repeatable (within one lab, one operator using one instrument or by one operator and one day) and reproducible (by different labs, different operators, on different days). The definition for reproducibility is similar to the one for repeatability, except reproducibility involves the variation of one or more of the following: lab, instrument, day, or operator (American Public Health Association, American Water Works Association, and Water Environment Federation. 1998. Standard methods for the examination of water and wastewater, 20th Ed. American Public Health Association, Washington, D.C.).**

> **Helpful document: for more details on the concepts of repeatability and reproducibility, see section 5.1 of latest proposed EPA guidance [EPA 2001. Guidance on Data Quality Indicators (EPA QA/G-5i) at http://on-linelearning.ca/idec4433/epaqaqc2000/g5i-prd.pdf.**
>
> **Field measurement probe makers seem to be equating repeatability or reproducibility with precision. Definitions vary somewhat among publications, but NIST definitions (see appendices) should be used for consistency.**

**When practicable, metadata should always include date, time, location, depth, flow, temperature and other default metadata fields suggested by the National Water Quality Monitoring Council and given a default place in EPA's new STORET database. To facilitate general data interpretation and PBMS data comparisons, STORET metadata for QC results (such as duplicate or triplicate precision comparisons) should include all field and lab precision and systematic error (bias) (QC sample) observed performance data AS WELL AS whatever summary statistics are calculated. In other words, if the QC measures for precision duplicate (repeatability) measures were 100 and 110, report 100 and 110 as part of metadata IN ADDITION TO (emphasis added) performance indicator summary statistics such as a NIST suggested standard deviation or a relative percent difference.**

> **Why is this so important? Because precision, systematic error (bias), and accuracy summary statistics are often given in different and incompatible units, making it hard combine them into total minimum measurement error or total minimum measurement uncertainty (see step IV-F.2).**
>
> **In STORET, include the type of sample (for example, type of reference sample or field spike or blank sample for systematic error (bias) in text boxes in the QC sample data entry form.**

**See appendix VI-A for more details.**

**VI-B. Data Reporting And Archiving:**

**For aquatic monitoring, the generic VS guidance (Outline for Vital Signs Monitoring Plans, 2003, http://science.nature.nps.gov/im/monitor/docs/monplan.doc) mentions the need to document special requirements for entering and managing water quality data in the Environmental Protection Agency's STORET database and mentions that NPS WRD has developed guidance to be followed. Accordingly, data reporting and archiving should follow the guidance provided in Part E (Guidance on Data Reporting and Archiving in STORET, http://science.nature.nps.gov/im/monitor/protocols/wqPartE.doc) as well as the more generalized NPS I&M guidance (science.nature.nps.gov/im/dmproto/joe40001.htm) as closely as possible within the practicalities of funding levels available.**

**A sustainable method for archiving and retrieving details of "protocol" method changes should be included in the plan.**

> **Note this is important because method changes can have serious consequences on comparability and thus on the ability to do long-term trend analyses. Thus, method changes should be kept to a minimum and a PBMS analysis should be done on the old and new methods to determine data comparability and adequacy for the task at hand.**

**Since STORET does not always provide sufficient room for all the metadata and other method details required for others to totally understand what was done and thus determine data usability and comparability, the plan needs to detail how the comments fields in STORET will be used for other method details, or alternative ways for such details to be stored and retrieved. For chemical studies involving both field work and then subsequent lab work, information to be included in the reporting packages, and stored for long term use and retrieval should typically include default EPA requirements (EPA. 1998. EPA Guidance For Quality Assurance Project Plans, EPA QA/G-5, EPA ORD**

**Publication No. EPA/600/R-98/018, http://www.epa.gov/quality/qs-docs/g5-final.pdf) such as:**

**Field Operation Records**

**Sample collection records.**
**Chain-of-custody records.**
**A Description of General field procedures.**
**Corrective action reports (if any)**

**Laboratory Records**

**Sample Data.**
**Sample Management Records.**
**Test Methods.**
**QA/QC (observed performance) Reports**

**Data Handling Records**

**Before being reported, data collected should go though a data validation and usability review. The plan should detail how reported values will be periodically compared to minimum and maximum plausible values.**

> **Note: As monitoring progresses, if observations that exceed minimum or maximum plausible values are recorded, additional calibration and other QA/QC steps should be taken to try to determine and eliminate the source of error. In any case, no values that exceed minimum or maximum values plausible (to the extent that professionals are absolutely sure such values are impossible) should be reported as part of final results. Typical short-term projects have a data validation step at the end of data collection. During long-term projects, data validation must be done repeatedly to make sure the data collection process has not gone awry. See definitions of useful quantitative data and useful qualitative data at the end of the appendices.**

**The plan should document the degree to which a subset of samples, particularly biological samples, will be archived in museums or other storage locations for potential future use.**

> **Why? Archiving of samples should be an essential component of a monitoring program given the likelihood of future improvements in analytical techniques and the development of new questions (M.L. Pace and J.J. Cole. 1989.What questions, systems or phenomena warrant long-term ecological study? In J. Likens, Ed. Long-Term Studies in Ecology, Approaches and Alternatives. Springer-Verlag, NY, p. 183-185).**

> **Older museum samples of fish, for example, have recently been analyzed for genetic fingerprints and for body burdens of contaminants, issues that were not thought about when the fish were originally collected.**

**See appendix VI-B for more details.**

## VII. DATA ANALYSIS AND REPORTING

**In concert with generic I&M vital signs guidance (Outline for Vital Signs Monitoring Plans, 2003, http://science.nature.nps.gov/im/monitor/docs/monplan.doc), this section of the plan should:**

- **Describe how data collected by the monitoring program will be analyzed, including who is responsible and how often analysis will occur.**
- **Describe the various reports and other products of the monitoring effort, including what they will include, who the intended audience is, how often they will be produced and in what format, and who will be responsible for ensuring that data are analyzed, reported, and sent to the right recipients in a timely manner.**

**Populations to be sampled, study units, and statistical analyses planned, were all first considered in step IV (above) and statistical "detection limits" were considered as part of protocol SOPs in step V (above). If the statistical analyses aspects are covered sufficiently in those parts of the monitoring plan and in attached protocols, there is no need to repeat the details here. In any case, who will do the analyses, and when they will be done, should be summarized in this section.**

## VIII. ADMINISTRATION/IMPLEMENTATION OF THE MONITORING PROGRAM

### VIII-A. General Documentation

**See generic Vital Signs guidance (Outline for Vital Signs Monitoring Plans, 2003, http://science.nature.nps.gov/im/monitor/docs/monplan.doc) for the basics of what is to be covered. As suggested in Vital Signs Monitoring Guidance, documentation in the plan of Administration/Implementation of the Monitoring Program shall include:**

- **Describe the makeup of the Board of Directors and Science/Technical committees for the network of parks, and their role in developing the monitoring strategy and implementing and promoting accountability for the monitoring program.**

- **What is the staffing plan for the monitoring program? Who will be involved in the program, where will be they be stationed, and what is their role in the program?**
- **Integration with other park operations: describe how the monitoring program integrates with other park operations such as interpretation, law enforcement, and maintenance.**
- **Partnerships: Describe other agencies and individuals that are part of the monitoring program. List cooperative agreements and other partnership agreements that are in place.**
- **For field sampling efforts to be performed in house, describe how they will be supported in terms of staff training and/or previous experience, field equipment to be dedicated to the effort (vehicles, instruments), anticipated in-house lab work to support operation, maintenance, and calibration of equipment and its documentation, and the necessary safety considerations in performing field tasks. (Note: each Network may want to standardize their own Safety Plan to cover monitoring efforts, particularly in regard to water quality sampling)**
- **Periodic Reviews: explain the process and schedule for periodic reviews of the overall program and various components and protocols.**

**In this section of the plan, documentation is provided that the proposed project management, budget, staff qualifications, and staff training, are all optimal (or at least reasonable) to ensure the success of the monitoring.**

## VIII-B. Project Management, Budget, Staff Qualifications, And Staff Training:

**See generic Vital Signs guidance (Outline for Vital Signs Monitoring Plans, 2003, http://science.nature.nps.gov/im/monitor/docs/monplan.doc) for the basics of what is to be covered**

**Typical NPS quality assurance and data quality objectives for aquatic projects:**

> **Descriptions of project management, project budget, monitoring staff qualifications, and monitoring staff training, should be included in the plan.**
>
> **The plan should include documentation that the level of project management commitment, time, and financial support line up with proposed monitoring activity and that the principle investigator should actively participate in the field data collection. The principle investigator should make sure that the training and motivation/enthusiasm of field crews remains high for the duration of the project in view of expected staff turnover, etc.**
>
>> **Note: a study of flawed/failed monitoring projects revealed that many of the problems could have been avoided if this had been done (L.M. Reid. 2001, The epidemiology of monitoring. Jour. Amer. Water Resources Assn. 37(4): 815-819).**

> **To help assure that data are of sufficient quality to be acceptable for various uses the plan therefore needs to document that the study staff will be properly managed and have the proper qualifications and training to produce quality data. Those chosen to do the monitoring and reporting need to have a good general reputation for scientific objectivity. Further, all data should be produced by a monitoring team whose project management, budget, staff qualifications, and staff training, are acceptable to the team as well as generally acceptable to suggested data users (regulatory or other).**

**A tabular summary of salaries, equipment, analytical, travel costs, specific dollar amounts for various categories of expenditure, and the total cost of the project is required. The total cost should be consistent with approved project cost.**

**To the extent practicable for long term monitoring, the plan should identify the name(s) of the principal project manager(s) investigators, and relevant outside specialists or consultants investigator(s) and their qualifications. Brief project-oriented résumés can be included as appendices to the plan.**

**All tasks, staffing needs, key personnel and organizations responsible for the completion of each task by milestone date (a schedule) need to be included in the plan.**

> **Note: specific tasks to address include persons responsible for environmental planning, coordination, and compliance needs; field sampling; preparation laboratory assignments; data management and archiving responsibilities; site preparation, and; construction monitoring.**

**In accordance with general I&M vital signs monitoring guidance (http://science.nature.nps.gov/im/monitor/monplan.doc), the plan should specifically provide detail on the administration/implementation of the Monitoring Program, including describing:**

- **The makeup of the Board of Directors and Science/Technical committees for the network of parks, and their role in developing the monitoring strategy and implementing and promoting accountability for the monitoring program.**
- **The staffing plan for the monitoring program. Who will be involved in the program, where will be they be stationed, and what is their role in the program?**
- **The integration with other park operations: describe how the monitoring program integrates with other park operations such as interpretation, law enforcement, and maintenance.**

- **Partnerships: Describe other agencies and individuals that are part of the monitoring program.**
- **Periodic Reviews: explain the process and schedule for periodic reviews of the overall program and various components and protocols.**
- **A description of how field sampling done in house will be supported in terms of staff training, dedicated field equipment (vehicles, instruments, etc.) lab work, Safety Plans, etc.)**

**The schedule should include anticipated delivery dates of progress and final report products and should include time for an adequate peer review of drafts.**

**For more details, see appendix VIII-B.**

## IX. SCHEDULE

**In concert with general Vital Signs guidance (Outline for Vital Signs Monitoring Plans, 2003, http://science.nature.nps.gov/im/monitor/docs/monplan.doc):**

**This part of the monitoring plan should summarize decisions on the proposed schedule and will begin to summarize the frequency of sampling (e.g., during what season of the year, and whether sampling should occur annually or once every several years) for the various components of the monitoring program. This part of the report should also summarize the target completion date for protocols still to be developed, or for other tasks that will require additional time to complete before a component of the monitoring will be implemented.**

## X. BUDGET

**This section of the plan will outline the budget for various tasks as detailed in generic Vital Signs Guidance (http://science.nature.nps.gov/im/monitor/monplan.doc).**

## XI. LITERATURE CITED

**As required in in generic Vital Signs Guidance (http://science.nature.nps.gov/im/monitor/monplan.doc).**

## APPENDICES:

**See Outline for Plans (op cit.):**

A. **Detailed descriptions of parks and their resources (optional)**

**B. Workshop reports**
**C. Sampling Protocols**
**D. Database design details**
**E. Data Management Plan**
**F. Results of Peer Review of Phase III (Including this in the final detailed study plan is optional, not required by the vital signs program in general. However, this would be a logical place to include it, and it is described here to help networks understand how this peer review step fits into the other general and QA/QC monitoring planning steps).**